

مركز البحوث

استكشاف البيانات

نظريات وخوارزميات وأمثلة

تأليف: د. نونغ يي

ترجمة

د. خالد بن ناصر آل حيان

راجع الترجمة

د. صالح بن محمد السليم


محمد خطاب

بسم الله الرحمن الرحيم



مركز البحوث



استكشاف البيانات

نظريات وخوارزميات وأمثلة

تأليف

د. نونغ يي

ترجمة

د. خالد بن ناصر آل حيان

راجع الترجمة

د. صالح بن محمد السليم

١٤٣٧هـ - ٢٠١٦م

بطاقة فهرسة

ⓗ معهد الإدارة العامة، ١٤٣٧هـ

فهرسة مكتبة الملك فهد الوطنية أثناء النشر

يي ، نونغ

استكشاف البيانات: نظريات وخوارزميات وأمثلة / نونغ يي؛ خالد بن ناصر آل حيان، صالح بن محمد السليم - الرياض ، ١٤٣٧هـ

٥٠٤ ص؛ ١٧ x ٢٤ س

ردمك: ٢-٢٤٤-١٤-

١- الخوارزمية (رياضيا

حيان، خالد بن ناصر

بن محمد (مراجع) ج- العنوان

ديوي: ٠٠٥,١٢ ١٤٣٧/٤٨٤٩

رقم الإيداع: ١٤٣٧/٤٨٤٩

ردمـــــــك: ٢-٢٤٤-١٤-٩٩٦٠

هذه ترجمة لكتاب

Data Mining
Theories, Algorithms, and Examples

ماتلاب (*MATLAB®*) هي علامة تجارية لشركة ماثووركس (*MathWorks*) ويتم استخدامها بتصريح. إن شركة ماثووركس غير مسئولة عن دقة النص أو التمارين الموجودة في هذا الكتاب. وإن استخدام هذا الكتاب أو البحث في برمجيات ماتلاب أو المنتجات ذات الصلة لا يشكل موافقة أو رعاية من قبل ماثووركس لنهج تعليمي معين أو استخدام معين لبرمجيات ماتلاب.

سي آر سي *(CRC)* للطباعة: مجموعة تايلور و فرانسيس (*TAYLOR &FRANCIS GROUP*) ٦٠٠ شارع بروكن ساوند باركواى شمال غرب ، الجناح ٣٠٠

مدينة بوكا راتون ، فلوريدا ٣٣٤٨٧-٢٧٤٢

سي ار سي للطباعة هي فرع من مجموعة تايلور وفرانسيس، مجموعة أعمال انفورما

ليس من حق أي جهة المطالبة بأعمال الحكومة الأمريكية الأصلية

رقم الكتاب المعياري الدولي ٢٠١٣٠٦٢٤ - المعايير الدولية للكتاب رقم : ٩٧٨-١-٤٣٩٨-٠٨٣٨-٢ (غلاف سميك)

http://www.taylorandfrancis.com

و موقع سي آر سي *(CRC)* للنشر
http://www.crcpress.com



جدول المحتويات

فهرس الجداول ١٢

فهرس الأشكـال ١٧

فهرس التمارين ٢١

تمهيد ٢٤

شكر وتقدير ٢٨

المؤلفة في سطور ٢٨

الجزء الأول: نظرة عامة على استكشاف البيانات ٢٩

١- مقدمة عن البيانات، وأنماط البيانات، واستكشاف البيانات ٣١

١-١ أمثلة عن مجموعات البيانات الصغيرة ٣١

١-٢ أنواع متغيرات البيانات ٣٦

١-٢-١ متغير الخاصية مقابل المتغير الهدف ٣٦

١-٢-٢ المتغير النوعي مقابل المتغير الرقمي ٤١

١-٣ أنماط البيانات التي يمكن استنباطها من خلال استكشاف البيانات ٤٢

١-٣-١ أنماط التصنيف والتنبؤ ٤٢

١-٣-٢ أنماط الاقتران وأنماط العنقود ٤٧

١-٣-٣ أنماط اختزال البيانات ٤٩

١-٣-٤ الأنماط المتطرفة والشاذة ٥١

١-٣-٥ الأنماط الزمنية والتسلسلية ٥٢

١-٤ البيانات التدريبية والبيانات الاختبارية ٥٤

التمارين ٥٥

الجزء الثاني: خوارزميات لاستكشاف أنماط التصنيف والتنبؤ ٥٧

٢- نماذج الانحدار الخطية وغير الخطية ٥٩

٢-١ نماذج الانحدار الخطي ٥٩

٢-٢ طريقة المربعات الصغرى وطريقة الإمكان الأكبر لتقدير المعلمة ٦٢

٢-٣ نماذج الانحدار غير الخطية وتقدير المعلمة ٦٩
٢-٤ البرمجيات والتطبيقات ٧١
التمارين ٧١
٣- مُصنّف بييز البسيط ٧٣
٣-١ نظرية بييز ٧٣
٣-٢ التصنيف القائم على نظرية بييز ومصنّف بييز البسيط ٧٣
٣-٣ البرمجيات والتطبيقات ٧٩
التمارين ٨٠
٤- أشجار القرار والانحدار ٨١
٤-١ تعلُّم شجرة القرار الثنائية وتصنيف البيانات باستخدام شجرة القرار ٨١
٤-١-١ عناصر شجرة القرار ٨١
٤-١-٢ شجرة القرار ذات طول الوصف الأصغر ٨٤
٤-١-٣ طرق انتقاء الانفصال ٨٥
٤-١-٤ خوارزمية بناء شجرة القرار من أعلى إلى أسفل ٩٢
٤-١-٥ تصنيف البيانات باستخدام شجرة القرار ١٠١
٤-٢ تعلُّم شجرة القرار غير الثنائية ١٠٣
٤-٣ التعامل مع القيم الرقمية والقيم المفقودة لمتغيرات الخاصية ١٠٩
٤-٤ التعامل مع متغير الهدف الرقمي وبناء شجرة الانحدار ١١٢
٤-٥ مزايا وعيوب خوارزمية شجرة القرار ١١٣
٤-٦ البرمجيات والتطبيقات ١١٨
التمارين ١١٩
٥- الشبكات العصبية الصناعية للتصنيف والتنبؤ ١٢١
٥-١ وحدات المعالجة للشبكات العصبية الصناعية ١٢١
٥-٢ معماريات الشبكات العصبية الصناعية ١٢٩

٥-٣ طرق تحديد أوزان الروابط في الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة ١٣٤
٥-٣-١ الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة (Perceptron) ١٣٤
٥-٣-٢ خصائص وحدة المعالجة ١٣٥
٥-٣-٣ الأسلوب البياني لتحديد أوزان الروابط والتحيزات ١٣٧
٥-٣-٤ طريقة تعلُم لتحديد أوزان الروابط والتحيزات ١٤٠
٥-٣-٥ عيوب الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة ١٤٤
٥-٤ طريقة التعلم بالتوالد الخلفي للشبكات العصبية الصناعية ذات التغذية الأمامية متعددة الطبقات ١٤٧
٥-٥ الاختيار التجريبي لمعمارية الشبكة العصبية الصناعية من أجل ملاءمة جيدة للبيانات ١٥٦
٥-٦ البرمجيات والتطبيقات ١٥٨
التمارين ١٥٨
٦- الدعم الآلي المتجه ١٦١
٦-١ الأساس النظري لصياغة وحل مشكلة التحسين لتعلُم دالة التصنيف ١٦١
٦-٢ صياغة الدعم الآلي المتجه (SVM) لمصنف خطي ولمشكلة قابلة للانفصال خطياً ١٦٣
٦-٣ التفسير الهندسي لصياغة الدعم الآلي المتجه (SVM) للمصنّف الخطي ١٧٠
٦-٤ حل المسألة البرمجية التربيعية لمصنّف خطي ١٧١
٦-٥ صياغة الدعم الآلي المتجه (SVM) لمصنّف خطي ولمسألة قابلة للفصل بشكل غير خطي ١٨٢
٦-٦ صياغة الدعم الآلي المتجه (SVM) لمصنّف غير خطي ومسألة قابلة للفصل بشكل غير خطي ١٨٦
٦-٧ طرق استخدام الدعم الآلي المتجه (SVM) لمسائل التصنيف متعددة الفئات ١٩٢

٦-٨ مقارنة بين الشبكة العصبية الصناعية (ANN) والدعم الآلي المتجه (SVM) ١٩٣
٦-٩ البرمجيات والتطبيقات ١٩٤
التمارين ١٩٥
٧- مصنّف أقربk - مجاور والتعنقُد المراقب ١٩٧
٧-١ مصنّف أقرب k-مجاور ١٩٧
٧-٢ التعنقد المراقب ٢٠٥
٧-٣ البرمجيات والتطبيقات ٢٢٥
التمارين ٢٢٥
الجزء الثالث: خوارزميات لاستكشاف أنماط العنقود والاقتران ٢٢٧
٨- التعنقد الهرمي ٢٢٧
٨-١ إجراء التعنقد الهرمي المحتشد ٢٢٩
٨-٢ طرق تحديد المسافة بين عنقودين ٢٣٠
٨-٣ توضيح كيفية إجراء التعنقد الهرمي ٢٣٧
٨-٤ الشجرة غير الرتيبة للتعنقد الهرمي ٢٤٣
٨-٥ البرمجيات والتطبيقات ٢٤٥
التمارين ٢٤٦
٩- التعنقد حول K- متوسط والتعنقد القائم على الكثافة ٢٤٧
٩-١ التعنقد حول K- متوسط ٢٤٧
٩-٢ التعنقد القائم على الكثافة ٢٦٣
٩-٣ البرمجيات والتطبيقات ٢٦٤
التمارين ٢٦٥
١٠- خريطة التنظيم الذاتي ٢٦٧
١٠-١ خوارزمية خريطة التنظيم الذاتي ٢٦٧
١٠-٢ البرامج والتطبيقات ٢٧٨

التمارين ٢٧٩

١١- التوزيعات الاحتمالية للبيانات الأحادية المتغير ٢٨١

١١-١ التوزيع الاحتمالي للبيانات الأحادية المتغير وخصائص التوزيع الاحتمالي لأنماط بيانات متنوعة ٢٨١

١١-٢ طريقة التمييز بين أربعة توزيعات احتمالية ٢٨٧

١١-٣ البرمجيات والتطبيقات ٢٨٩

التمارين ٢٩٠

١٢- قواعد الاقتران ٢٩١

١٢-١ تعريف قواعد الاقتران ومقاييس الاقتران ٢٩١

١٢-٢ اكتشاف قاعدة الاقتران ٢٩٧

١٢-٣ البرمجيات والتطبيقات ٣٠٥

التمارين ٣٠٦

١٣- شبكة بييز ٣٠٧

١٣-١ بُنية شبكة بييز والتوزيعات الاحتمالية للمتغيرات ٣٠٧

١٣-٢ الاستدلال الاحتمالي ٣١٧

١٣-٣ تعلُّم شبكة بييز ٣٢٦

١٣-٤ البرمجيات والتطبيقات ٣٢٩

التمارين ٣٢٩

الجزء الرابع: خوارزميات استكشاف أنماط اختزال البيانات ٣٣١

١٤- تحليل المكونات الرئيسية ٣٣٣

١٤-١ مراجعة لإحصاءات المتغيرات المتعددة ٣٣٣

١٤-٢ مراجعة جبر المصفوفات ٣٣٨

١٤-٣ تحليل المكونات الرئيسة ٣٤٩

١٤-٤ البرمجيات والتطبيقات ٣٥٣

التمارين ٣٥٣

١٥- القياس المتعدد الأبعاد ٣٥٥
١٥-١ خوارزمية القياس المتعدد الأبعاد ٣٥٥
١٥-٢ عدد الأبعاد ٣٧٧
١٥-٣ قياس الفروقات الفردية للقياس المتعدد الأبعاد الموزون ٣٧٧
١٥-٤ البرمجيات والتطبيقات ٣٧٨
التمارين ٣٧٩
الجزء الخامس: خوارزميات استكشاف الأنماط المتطرفة والشاذة ٣٨١
١٦- مخطط التحكم أحادي المتغير ٣٨٣
١٦-١ مخططات التحكم لشوارتز ٣٨٣
١٦-٢ مخططات تحكم المجموع التراكمي ٣٨٨
١٦-٣ مخططات التحكم للمتوسط المتحرك الموزون الأسّي ٣٩٣
١٦-٤ مخططات تحكم الدرجة التراكمية ٣٩٩
١٦-٥ منحنى التشغيل التشخيصي لتقييم ومقارنة مخططات التحكم ٤٠٤
١٦-٦ البرمجيات والتطبيقات ٤٠٨
التمارين ٤٠٨
١٧- مخططات التحكم متعددة المتغيرات ٤١١
١٧-١ مخططات التحكم لهوتلينق T2 .... ٤١١
١٧-٢ مخططات تحكم المتوسط المتحرك الموزون الأسي متعددة المتغيرات ٤١٥
١٧-٣ مخططات تحكم مربع كاي ٤١٦
١٧-٤ التطبيقات ٤١٨
التمارين ٤١٩
الجزء السادس: خوارزميات استكشاف الأنماط الزمنية والتسلسلية ٤٢١
١٨- تحليل الارتباط الذاتي والسلاسل الزمنية ٤٢٣
١٨-١ الارتباط الذاتي ٤٢٣
١٨-٢ السكون واللاسكون ٤٢٥

١٨-٣ نماذج المتوسط المتحرك ذاتي الانحدار الخاصة ببيانات السلاسل الساكنة ٤٢٦
١٨- ٤ خصائص دالة الارتباط الذاتي ودالة الارتباط الذاتي الجزئي لنماذج المتوسط المتحرك ذاتي الانحدار ٤٣٠
١٨-٥ تحويل بيانات السلسلة غير الساكنة ونماذج المتوسط المتحرك المتكامل ذاتي الانحدار ٤٣٢
١٨-٦ البرمجيات والتطبيقات ٤٣٤
التمارين ٤٣٥
١٩- نماذج سلسلة ماركوف ونماذج ماركوف المخفية ٤٣٧
١٩-١ نماذج سلسلة ماركوف ٤٣٧
١٩-٢ نماذج ماركوف المخفية ٤٤٢
١٩-٣ تعلم نماذج ماركوف المخفية ٤٤٧
١٩-٤ البرمجيات والتطبيقات ٤٦٢
التمارين ٤٦٢
٢٠- تحليل المويجة ٤٦٣
٢٠-١ تعريف المويجة ٤٦٣
٢٠-٢ تحويل المويجة لبيانات السلاسل الزمنية ٤٦٥
٢٠-٣ إعادة بناء السلسلة الزمنية الزمن من معاملات المويجة ٤٧٦
٢٠-٤ البرمجيات والتطبيقات ٤٧٨
التمارين ٤٧٩
المراجع – References ٤٨١
قاموس المصطلحات - Glossary ٤٨٩

فهرس الجداول

الصفحة	الجدول
٣٣	الجدول ١-١: مجموعة بيانات البالون
٣٥	الجدول ١-٢: مجموعة البيانات الخاصة بالحلقات الدائرية في مكوك الفضاء
٣٧	الجدول ١-٣: مجموعة البيانات الخاصة بالعدسات
٣٩	الجدول ١-٤: مجموعة البيانات الخاصة باكتشاف الأعطال وتشخيصها في نظام تصنيع معين
٤٦	الجدول ١-٥: القيمة المتوقعة لعدد الحلقات الدائرية ذات الأحمال الثقيلة
٥٣	الجدول ١-٦: مجموعة بيانات اختبارية لنظام تصنيع معين لاكتشاف وتشخيص الأعطال
٦٧	الجدول ٢-١: مجموعة بيانات الحلقات الدائرية ذات الأحمال الثقيلة مع القيمة المستهدفة المتوقعة من الانحدار الخطي
٦٩	الجدول ٢-٢: العملية الحسابية لتقدير معلمات النموذج الخطي في المثال ٢-١
٧٦	الجدول ٣-١: مجموعة البيانات التدريبية الخاصة بالكشف عن أعطال نظام التصنيع
٧٧	الجدول ٣-٢: تصنيف سجلات البيانات في مجموعة البيانات التدريبية الخاصة بالكشف عن أعطال نظام التصنيع
٨٢	الجدول ٤-١: مجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع
٨٧	الجدول ٤-٢: الانفصال الثنائي لعقدة الجذر والعملية الحسابية لقيمة مقياس عشوائية المعلومات لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع
٩٤	الجدول ٤-٣: الانفصال الثنائي لعقدة الجذر والعملية الحسابية لقيمة مؤشر جيني لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع
٩٧	الجدول ٤-٤: الانقسام الثنائي للعقدة الداخلية مع D={2,4,5,9,10}، وحساب مقياس عشوائية المعلومات لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع

الصفحة	الجدول
٩٩	الجدول ٤-٥: الانقسام الثنائي للعقدة الداخلية المحتوية على D={2,4,5,9,10}، وحساب مؤشر جيني لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع
١٠٢	الجدول ٤-٦: تصنيف سجلات البيانات لمجموعة البيانات الاختبارية الخاصة بالكشف عن أعطال نظام التصنيع
١٠٧	الجدول ٤-٧: الانفصال غير الثنائي لعقدة الجذر وعملية حساب مقياس عشوائية المعلومات لمجموعة بيانات العدسات
١٠٨	الجدول ٤-٨: الانفصال غير الثنائي للعقدة الداخلية {2، 4، 6، 8، 10، 12، 14، 16، 18، 20، 22، 24}، وعملية حساب مقياس عشوائية المعلومات لمجموعة بيانات العدسات
١٠٩	الجدول ٤-٩: الانفصال غير الثنائي للعقدة الداخلية {2، 6، 10، 14، 18، 22}، وعملية حساب مقياس عشوائية المعلومات لمجموعة بيانات العدسات.
١١٠	الجدول ٤-١٠: الانفصال غير الثنائي للعقدة الداخلية {4، 8، 12، 16، 20، 24} وعملية حساب مقياس عشوائية المعلومات لمجموعة بيانات العدسات.
١٢٦	الجدول ٥-١: الدالة ***AND***
١٢٨	الجدول ٥-٢: الدالة ***OR***
١٣٣	الجدول ٥-٣: الدالة ***XOR***
١٤٦	الجدول ٥-٤: دالة خاصة بكل وحدة معالجة في شبكة الـ ***ANN*** ثنائية الطبقات لتطبيق الدالة ***XOR***
١٤٦	الجدول ٥-٥: الدالة ***NOT***
١٧٧	الجدول ٦-١: الدالة ***AND***
٢٠٣	الجدول ٧-١: مجموعة البيانات التدريبية الخاصة بالكشف عن الأعطال بنظام التصنيع
٢٠٤	الجدول ٧-٢: مجموعة البيانات الاختيارية الخاصة بالكشف عن الأعطال بنظام التصنيع ونتائج التصنيف في الأمثلة ٧-١ و٧-٢

الجدول	الصفحة
الجدول ٧-٣: خوارزمية التعنقد المراقب - (إنجليزي وعربي)	٢٠٩
الجدول ٨-١: مجموعة البيانات الخاصة باكتشاف أعطال النظام مع تسع حالات من الأعطال الآلية الأحادية	٢٣٧
الجدول ٨-٢: المسافة لكل زوج من العناقيد: C_1، C_2، C_3، C_4، C_5، C_6، C_7، C_8، و C_9	٢٣٩
الجدول ٨-٣: مسافة كل زوج من العناقيد: $C_{1,5}$، $C_{2,4}$، C_3، $C_{6,7}$، C_8 و C_9	٢٤٠
الجدول ٨-٤: مسافة كل زوج من العناقيد: $C_{1,5}$، $C_{2,4,8}$، C_3، $C_{6,7}$، و C_9	٢٤١
الجدول ٨-٥ :مسافة كل زوج من العناقيد: $C_{1,5,6,7,9}$، $C_{2,4,8}$، و C_3	٢٤٢
الجدول ٩-١: خوارزمية التعنقد حول ***K***-متوسط - (إنجليزي وعربي)	٢٤٨
الجدول ٩-٢: مجموعة البيانات لاكتشاف أعطال النظام بتسع حالات من الأعطال الآلية الأحادية	٢٥٠
الجدول ١٠-١: خوارزمية التعلم لخريطة التنظيم الذاتي ***(SOM)*** - (إنجليزي وعربي)	٢٧٠
الجدول ١٠-٢: مجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع بتسع حالات للأعطال الآلية الأحادية	٢٧٢
الجدول ١١-١: قيم درجة حرارة الإطلاق (*Launch Temperature*) في مجموعة البيانات الخاصة بعدد الحلقات الدائرية في مكوك الفضاء	٢٨٢
الجدول ١١-٢: خليط من نتائج اختبارات الانحراف *(Skewness)* والنسق *(Mode)* لتمييز التوزيعات الاحتمالية الأربعة	٢٨٨
الجدول ١٢-١: مجموعة بيانات اكتشاف أعطال النظام بتسع حالات من الأعطال الآلية الأحادية ومجموعات العنصر التي تمّ الحصول عليها من مجموعة البيانات هذه	٢٩٢
الجدول ١٢-٢: خوارزمية أبريوري (الأسبقية) (*Apriori Algorithm*) - (إنجليزي وعربي)	٢٩٩
الجدول ١٣-١: مجموعة البيانات التدريبية الخاصة باكتشاف أعطال نظام تصنيع	٣٠٨
الجدول ١٣-٢: إيجاد احتمال $P(x_5 \| x_1)$	٣١١

الصفحة	الجدول
٣١١	الجدول ١٣-٣: إيجاد احتمال $P(x_6\| x_3)$
٣١١	الجدول ١٣-٤: إيجاد احتمال $P(x_4\| x_3, x_2)$
٣١٢	الجدول ١٣-٥: إيجاد احتمال $P(x_9\| x_5)$
٣١٢	الجدول ١٣-٦: إيجاد احتمال $P(x_7\| x_5, x_6)$
٣١٢	الجدول ١٣-٧: إيجاد احتمال $P(x_8\| x_4)$
٣١٣	الجدول ١٣-٨: إيجاد احتمال $P(y\| x_9)$
٣١٣	الجدول ١٣-٩: إيجاد احتمال $P(y\| x_7)$
٣١٣	الجدول ١٣-١٠: إيجاد احتمال $P(y\| x_8)$
٣١٤	الجدول ١٣-١١: إيجاد احتمال $P(x_1)$
٣١٤	الجدول ١٣-١٢: إيجاد احتمال $P(x_2)$
٣١٤	الجدول ١٣-١٣: إيجاد احتمال $P(x_3)$
٣٣٧	الجدول ١٤-١: مجموعة البيانات الخاصة بالكشف عن الأعطال بنظام التصنيع مع متغيرين للجودة
٣٣٧	الجدول ١٤-٢: الاحتمالات المشتركة والهامشية لمتغيري الجودة
٣٥٧	الجدول ١٥-١: خوارزمية القياس المتعدد الأبعاد (*MDS*) - (إنجليزي وعربي)
٣٦٠	الجدول ١٥-٢: خوارزمية الاتحاد الرتيبة - (إنجليزي وعربي)
٣٦٥	الجدول ١٥-٣: مجموعة البيانات لنظام اكتشاف الأعطال مع ثلاث حالات من الأعطال الآلية الأحادية
٣٦٥	الجدول ١٥-٤: المسافة الإقليدية لكل زوج من سجلات البيانات
٣٨٤	الجدول ١٦-١: عينات من ملحوظات البيانات المرصودة
٣٩١	الجدول ١٦-٢: ملحوظات البيانات المرصودة لدرجة حرارة الإطلاق من مجموعة بيانات الحلقات الدائرية ذات الأحمال الثقيلة جنباً إلى جنب مع الإحصائيات لمخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب
٣٩٦	الجدول ١٦-٣: ملحوظات البيانات المرصودة لدرجة حرارة الإطلاق مجموعة بيانات الحلقات الدائرية ذات الأحمال الثقيلة جنباً إلى جنب مع إحصائية *EWMA* لمخطط تحكم الـ *EWMA*

الجدول	الصفحة
الجدول ١٦-٤ : أزواج من معدل الإنذار الخاطئ ومعدل الزيارة الناجحة لقيم متنوعة من حد القرار H لمخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب في المثال ١٦-١	٤٠٦
الجدول ١٧-١: مجموعة البيانات لاكتشاف أعطال النظام مع اثنين من متغيرات الجودة x_7 و x_8	٤١٥
الجدول ١٨-١: سلسلة زمنية لنموذج الانحدار الذاتي $AR(1)$ حيث $\phi_1 = 0.09$، $x_0 = 3$ وخطأ عشوائي e_t	٤٢٧
الجدول ١٨-٢: سلسلة زمنية لنموذج $MA(1)$ مع $\theta_1 = 0.9$ وخطأ عشوائي e_t	٤٢٩

فهرس الأشكال

الصفحة	الشكل
٤٠	الشكل ١-١: خريطة نظام تصنيع معين ذو تسع آلات وتدفقات إنتاج وحدات المنتج
٤٤	الشكل ١-٢: النموذج الملائم للعلاقة الخطية الخاصة بدرجة حرارة الإطلاق مع عدد الحلقات الدائرية ذات الأحمال الثقيلة في مجموعة البيانات الخاصة بالحلقات الدائرية في مكوك الفضاء
٤٨	الشكل ١-٣: التعنقد الخاص بـ 10 سجلات من سجلات البيانات في مجموعة بيانات نظام التصنيع
٥٠	الشكل ١-٤: اختزال البيانات ثنائية الأبعاد إلى مجموعة من البيانات ذات بعد واحد
٥١	الشكل ١-٥: الرسم البياني التكراري لدرجات حرارة الإطلاق في مجموعة بيانات مكوك الفضاء
٥٢	الشكل ١-٦ :درجة حرارة الطقس كل ثلاثة شهور لمدة ٣ سنوات
٦٠	الشكل ٢-١: مثال توضيحي لنموذج انحدار بسيط
٨٣	الشكل ٤-١: شجرة القرار الخاصة بالكشف عن أعطال نظام التصنيع
٩٠	الشكل ٤-٢: عشوائية المعلومات
١٠١	الشكل ٤-٣: تصنيف سجل بيانات بدون عطل نظام باستخدام شجرة القرار الخاصة بالكشف عن أعطال نظام التصنيع
١٠٣	الشكل ٤-٤: تصنيف سجل بيانات لأعطال متعددة الآلات باستخدام شجرة قرار خاصة بالكشف عن أعطال نظام التصنيع
١٠٦	الشكل ٤-٥: شجرة القرار لمجموعة بيانات العدسات
١١٦	الشكل ٤-٦: شجرة القرار لمجموعة البيانات الخاصة بالبالون
١٢٢	الشكل ٥-١: وحدة معالجة بالشبكة العصبية الصناعية (*ANN*)
١٢٤	الشكل ٥-٢: أمثلة على دوال التحول
١٢٥	الشكل ٥-٣: تطبيق الدالة ***AND*** باستخدام وحدة معالجة واحدة
١٢٩	الشكل ٥-٤: تطبيق الدالة ***OR*** باستخدام وحدة معالجة واحدة

الصفحة	الشكل
١٣٠	الشكل ٥-٥: معمارية الشبكات العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة
١٣١	الشكل ٥-٦: معمارية الشبكات العصبية الصناعية ذات التغذية الأمامية ثنائية الطبقات
١٣٢	الشكل ٥-٧: شبكات عصبية صناعية ذات تغذية أمامية ثنائية الطبقات تطبق دالة ***XOR***
١٣٣	الشكل ٥-٨: معماريات الشبكات العصبية الصناعية الدورية
١٣٦	الشكل ٥-٩: مثال على حد القرار وفصل بين فضاء المدخلات إلى منطقتين من خلال وحدة المعالجة
١٣٨	الشكل ٥-١٠: توضيح الطريقة البيانية لتحديد أوزان الروابط
١٤١	الشكل ٥-١١: توضيح طريقة تعلم تغيير أوزان الروابط
١٤٦	الشكل ٥-١٢: نقاط البيانات الأربع للدالة ***XOR***
١٥١	الشكل ٥-١٣: مجموعة من الأوزان بقيم عشوائية في شبكة الـ *ANN* ذات التغذية الأمامية ثنائية الطبقات للدالة *XOR*
١٥٥	الشكل ٥-١٤: أثر معدل التعلم
١٥٧	الشكل ٥-١٥: مثال يوضح نموذجاً غير خطي مفرط في مطابقة البيانات من نموذج خطي
١٦٦	الشكل ٦-١: الدعم الآلي المتجه (*SVM*) لمصنّف خطي ومشكلة قابلة للانفصال خطيا. (*a*) حد القرار ذو هامش كبير. (*b*) حد القرار ذو هامش صغير
١٨١	الشكل ٦-٢: دالة القرار ومتجهات الدعم للمصنّف الخطي الخاص بالدعم الآلي المتجه *SVM* في المثال ٦-١
١٩١	الشكل ٦-٣: دالة قرار كثيرة الحدود في فضاء ثنائي الأبعاد
١٩٢	الشكل ٦-٤: دالة قاعدة دائرية لقوسشيان في فضاء ثنائي الأبعاد
٢٣٩	الشكل ٨-١: نتيجة التعنقد الهرمي لمجموعة بيانات اكتشاف أعطال النظام
٢٤٤	الشكل ٨-٢: مثال على ثلاث نقاط بيانات والتي تنتج لها طريقة ترابط المركز المتوسط شجرة غير رئيسية للتعنقد الهرمي

الصفحة	الشكل
٢٤٥	الشكل ٨-٣: الشجرة غير الرئيسية للتعنقد الهرمي لنقاط البيانات في الشكل ٨-٢
٢٦٨	الشكل ١٠-١: التصاميم الخاصة بخريطة التنظيم الذاتي (*SOM*) بخريطة مخرجات (a) أحادية، (b) ثنائية، و (c) وثلاثية الأبعاد
٢٧٢	الشكل ١٠-٢: التصاميم الخاصة بخريطة التنظيم الذاتي (*SOM*) للمثال ١٠-١
٢٧٥	الشكل ١٠-٣: العقد الفائزة لنقاط البيانات التسع في المثال ١٠-١ باستخدام قيم الوزن أولية
٢٨٣	الشكل ١١-١: المدرج التكراري لبيانات درجة حرارة الإطلاق (*Launch Temperature*)
٢٨٥	الشكل ١١-٢: أنماط بيانات السلاسل الزمنية وتوزيعاتها الاحتمالية
٢٩٦	الشكل ١٢-١: نظام تصنيع يحتوي على تسع آلات وخط إنتاج وحدات المنتج
٣١٠	الشكل ١٣-١: نظام تصنيع بتسع آلات وتدفقات إنتاج لوحدات المنتج
٣١٠	الشكل ١٣-٢: البنية (structure) الخاصة بشبكة بييز لمجموعة بيانات اكتشاف أعطال نظام التصنيع
٣٤٠	الشكل ١٤-١: حساب طول المتجه
٣٤٠	الشكل ١٤-٢: حساب الزاوية بين متجهين
٣٧٦	الشكل ١٥-١: مثال على رسم الجهد الخاص بنتيجة القياس المتعدد الأبعاد (*MDS*)مقابل عدد الأبعاد
٣٩٢	الشكل ١٦-١: مخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب لدرجة حرارة الإطلاق في مجموعة بيانات الحلقة الدائرية ذات الأحمال الثقيلة
٣٩٥	الشكل ١٦-٢: أوزان متناقصة أسيا على ملحوظات البيانات المرصودة
٣٩٧	الشكل ١٦-٣: مخطط تحكم *EWMA* لمراقبة درجة حرارة الإطلاق من مجموعة بيانات الحلقات الدائرية ذات الأحمال الثقيلة
٤٠٧	الشكل ١٦-٤: منحنى التشغيل التشخيصي (*ROC*) لمخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب في المثال ١٦-١

الصفحة	الشكل
٤١٢	الشكل ١٧-١: توضيح للمسافة الإحصائية المُقاسة باستخدام إحصاءة هوتلينق T^2 وحدود التحكم لمخططات التحكم لهوتلينق T^2 ومخططات التحكم أحادية المتغير
٤٢٧	الشكل ١٨-١: بيانات سلسلة زمنية يتم توليدها باستخدام نموذج الانحدار الذاتي $AR(1)$ حيث $\phi_1 = 0.09$، و $x_0 = 3$ وخطأ عشوائي e_t
٤٢٩	الشكل ١٨-٢: بيانات سلسلة زمنية تم توليدها باستخدام نموذج $MA(1)$ مع $\theta_1 = 0.9$ وخطأ عشوائي e_t
٤٤٠	الشكل ١٩-١: الحالات وانتقال الحالات في المثال ١٩-١
٤٤٣	الشكل ١٩-٢: أي طريقة من طرق المسار وطريقة المسار الأفضل لنماذج ماركوف المخفية
٤٦٤	الشكل ٢٠-١: دالة القياس ودالة المويجة لمويجة هار وآثار التمدد (*Dilation*) والتحويل (*Shift*)
٤٦٦	الشكل ٢٠-٢: عينة من بيانات سلسلة زمنية من (*a*) دالة، (*b*) عينة من سجلات البيانات مأخوذة من الدالة، و(*c*) تقريب الدالة باستخدام دالة القياس لمويجة هار
٤٧٤	الشكل ٢٠-٣: توضيح بياني لمويجة باول، ومويجة (*DoG*) اشتقاق مويجة قوسشيان، ومويجة داوبيشيز، ومويجة مورليت. (يي، إن، نظم الحاسوب والشبكة الآمنة: النمذجة والتحليل والتصميم، ٢٠٠٨، الشكل ١١,٢، ص ٢٠٠ حقوق الطبع والنشر لشركة وايلي في سي اتش فيرلاغ وشركاه المحدودة) - (*Ye, N., Secure Computer and Network Systems: Modeling, Analysis and Design, 2008, Figure 11.2, p. 200. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission*).

فهرس التمارين

التمارين	الصفحة
تمارين الفصل الأول (مقدمة عن البيانات، وأنماط البيانات، واستكشاف أنماط البيانات)	٥٥
تمارين الفصل الثاني (نماذج الانحدار الخطية وغير الخطية)	٧١
تمارين الفصل الثالث (مُصنّف بييز البسيط)	٨٠
تمارين الفصل الرابع (أشجار القرار والانحدار)	١١٩
تمارين الفصل الخامس (الشبكات العصبية الصناعية للتصنيف والتنبؤ)	١٥٨
تمارين الفصل السادس (الدعم الآلي المتجه)	١٩٥
تمارين الفصل السابع (مصنّف أقرب k - مجاور والتعنقُد المراقب)	٢٢٥
تمارين الفصل الثامن (التعنقد الهرمي)	٢٤٦
تمارين الفصل التاسع (التعنقد حول K- متوسط والتعنقد القائم على الكثافة)	٢٦٥
تمارين الفصل العاشر (خريطة التنظيم الذاتي)	٢٧٩
تمارين الفصل الحادي عشر (التوزيعات الاحتمالية للبيانات أحادية المتغير)	٢٩٠
تمارين الفصل الثاني عشر (قواعد الاقتران)	٣٠٦
تمارين الفصل الثالث عشر (شبكة بييز)	٣٢٩
تمارين الفصل الرابع عشر (تحليل المكونات الرئيسية)	٣٥٣
تمارين الفصل الخامس عشر (القياس المتعدد الأبعاد)	٣٧٩
تمارين الفصل السادس عشر (مخطط التحكم أحادي المتغير)	٤٠٨
تمارين الفصل السابع عشر (مخططات التحكم متعددة المتغيرات)	٤١٩
تمارين الفصل الثامن عشر (تحليل الارتباط الذاتي والسلاسل الزمنية)	٤٣٥
تمارين الفصل التاسع عشر (نماذج سلسلة ماركوف ونماذج ماركوف المخفية)	٤٦٢
تمارين الفصل العشرين (تحليل المويجة)	٤٧٩

تمهيد:

لقد مكنتنا التقنيات الحديثة من جمع كميات هائلة من البيانات في العديد من المجالات. وعلى الرغم من ذلك فإن سرعتنا في اكتشاف معلومات ومعرفة مفيدة من هذه البيانات أقل بكثير من سرعتنا في جمع تلك البيانات. وتستلزم عملية تحويل كَم هائل من البيانات إلى معلومات ومعرفة مفيدة القيام بخطوتين، هما: (١) البحث والتنقيب عن الأنماط التي تتخذها تلك البيانات و(٢) تفسير أنماط البيانات تلك ضمن نطاق المشكلة المستهدفة لتحويل هذه الأنماط إلى معلومات ومعرفة مفيدة.

يوجد العديد من خَوارِزْميّات استكشاف البيانات لغرض أتمتة الخطوة الأولى الخاصة بالبحث عن أنماط بيانات متنوعة في كم هائل من البيانات. وعادةً ما يعتمد تفسير أنماط البيانات المكتشفة على المعرفة بنطاق المشكلة المستهدفة إضافةً إلى القدرة على التفكير التحليلي. ويتناول هذا الكتاب التعرف على خَوارِزْميّات الاستكشاف والتنقيب عن البيانات التي يمكن استخدامها في استكشاف أنواع مختلفةَ من أنماط البيانات. وسوف يمكننا تعلُّم وتطبيق خُوارِزْميّات استكشاف البيانات من أتمتة ومن ثمّ تسريع عملية تنفيذ الخطوة الأولى الخاصة بالكشف عن أنماط البيانات من كمٍ هائل من البيانات. إن معرفة كيفية استنباط أنماط البيانات بواسطة تلك الخوارزميات يُعد أمراً شديد الأهمية لتنفيذ الخطوة الثانية ألا وهى تحديد معنى أنماط البيانات ضمن نطاق المشكلة النابعة منها ومن ثمّ تحويل أنماط تلك البيانات إلى معلومات ومعارف مفيدة.

نبذة عن الكتاب:

تمّ تنظيم خوارزميات استكشاف البيانات في هذا الكتاب ضمن خمسة أجزاء، كل جزء منه يستعرض كيفية الاستكشاف عن أحد أنواع أنماط البيانات الخمسة من كم هائل من البيانات، وهذه الأنماط هي كما يلي:

١) أنماط التصنيف والتنبؤ
٢) أنماط الاقتران وأنماط العنقود
٣) أنماط اختزال البيانات
٤) الأنماط المتطرفة والشاذة
٥) الأنماط الزمنية والتسلسلية

يستعرض الجزء الأول من الكتاب هذه الأنواع من أنماط البيانات مع ذكر أمثلة توضيحية. أما الأجزاء الخمسة الباقية من الكتاب - بدايةً من الجزء الثاني وحتى الجزء السادس - فقد عُنيت بوصف خوارزميات استكشاف الأنواع الخمسة من أنماط البيانات على التوالي.

وتركز أنماط التصنيف والتنبؤ على العلاقة بين متغيرات الخاصية ومتغيرات الهدف، وهو ما يسمح لنا بتصنيف أو التنبؤ بقيم متغيرات الهدف بناءً على قيم متغيرات الخاصية. ويتناول الجزء الثاني من الكتاب الخوارزميات التالية والتي تُستخدم في استكشاف أنماط التصنيف والتنبؤ:

- نماذج الانحدار الخطية وغير الخطية (الفصل ٢)
- مصنِّف بييز البسيط (الفصل ٣)
- أشجار القرار والانحدار (الفصل ٤)
- الشبكات العصبية الصناعية (*Artificial Neural Networks - ANNs*) للتصنيف والتنبؤ (الفصل ٥)
- الدعم الآلي المتجه (*Support Vector Machines - SVM*) (الفصل ٦)
- مصنِّف أقرب k - مجاور والتعنقد المراقب (الفصل ٧)

في حين يصف الجزء الثالث من الكتاب خوارزميات استكشاف البيانات المستخدمة لاستنباط أنماط الاقتران وأنماط العنقود. حيث تكشف أنماط العنقود عن أوجه التشابه والاختلاف بين سجلات البيانات. ويتم استنباط أنماط الاقتران على أساس التلازم في حدوث العناصر الموجودة في سجلات البيانات. باختصار، يصف الجزء الثالث خوارزميات استكشاف البيانات التالية للبحث عن أنماط الاقتران وأنماط العنقود:

- التعنقُد الهرمي (الفصل ٨)
- التعنقد حول K من المتوسطات والتعنقد على أساس الكثافة (الفصل ٩)
- خريطة التنظيم الذاتي (الفصل ١٠)
- التوزيعات الاحتمالية للبيانات أحادية المتغير (الفصل ١١)
- قواعد الاقتران (الفصل ١٢)
- شبكات بييز (الفصل ١٣)

أمـا أنمـاط اختزال البيـانـات، فهي تبحـث عن عـدد قليـل من المتغيرات التي يمكن اســـتخدامها لتمثيل مجموعة من البيانات ذات عدد أكبر بكثير من المتغيرات. وحيث إن المتغير الواحد يعطي بعداً واحداً من البيانات، فإن أنماط اختزال البيانات تســـمح بتمثيل مجموعة من البيانات موجودة في فضــاء متعدد الأبعاد في فضـــاء أقل من الأبعاد. يصــف الجزء الرابع خوارزميات استكشاف البيانات التالية للبحث عن أنماط اختزال البيانات:

- تحليل المكونات الرئيسية (الفصل ١٤)
- القياس المتعدد الأبعاد (الفصل ١٥)

وبالنسبة للقيم المتطرفة والشاذة، فهي نقاط البيانات التي تختلف بشكل كبير عن التعريف العادي والمعياري للبيانات، وهناك طرق عديدة لتعريف وإنشاء التعريف المعياري للبيانات. يصف الجزء الخامس خوارزميات استكشاف البيانات التالية لكشف وتحديد القيم المتطرفة والشاذة:

- مخطط التحكم أحادي المتغير (الفصل ١٦)
- مخطط التحكم متعدد المتغيرات (الفصل ١٧)

من ناحية أخرى، تكشف الأنماط الزمنية والتسلسلية كيفية تغير أنماط البيانات على مر الزمن. ويصف الجزء السادس خوارزميات استكشاف البيانات التالية للبحث عن الأنماط التسلسلية والزمنية:

- تحليل الارتباط الذاتي وسلاسل الزمن (الفصل ١٨)
- نماذج سلسلة ماركوف ونماذج ماركوف المخفية (الفصل ١٩)
- تحليل المويجات (الفصل ٢٠)

المزايا الرئيسة لهذا الكتاب:

كما أوضحنا سابقاً، تُعدُ عملية الاستكشاف والتنقيب عن أنماط البيانات في كم هائل من البيانات هي فقط الخطوة الأولى لتحويل البيانات إلى معلومات ومعرفة مفيدة ضمن نطاق المشكلة المستهدفة. ويجب أن يتم فهم وتفسير أنماط البيانات ضمن نطاق المشكلة الخاصة بها من أجل أن تكون مفيدة وذات معنى. ولتطبيق خوارزمية استكشاف البيانات

والتمكن من فهم وتفسير أنماط البيانات الناتجة من تطبيق الخوارزمية، نحتاج إلى فهم جانبين مهمين من الخوارزمية:

١) المفاهيم النظرية التي ترسّخ الأساس المنطقي لتبرير وضع عناصر خوارزمية استكشاف البيانات معاً بطريقة محددة للبحث عن نوع معين من نمط البيانات.

٢) الخطوات التشغيلية والتفاصيل الخاصة بكيفية معالجة خوارزمية استكشاف البيانات لكمٍّ هائل من البيانات من أجل الحصول على أنماط البيانات.

يهدف هذا الكتاب إلى تقديم كل من المفاهيم النظرية والتفاصيل التشغيلية لخوارزميات استكشاف البيانات في كل فصل بطريقة قائمة بذاتها ومتكاملة مع إعطاء أمثلة من البيانات الصغيرة. مما سيعمل على تمكين القارئ من فهم الجوانب النظرية والعملية لخوارزميات استكشاف البيانات، وتنفيذ الخوارزميات يدوياً من أجل الوصول إلى فهم شامل لأنماط البيانات الناتجة عن الخوارزميات.

يغطي هذا الكتاب خوارزميات استكشاف البيانات الموجودة بشكل شائع في الدراسات والمؤلفات الخاصة باستكشاف البيانات (على سبيل المثال، خوارزمية أشجار القرار، وخوارزمية الشبكات العصبية الصناعية، وخوارزمية التعنقُد الهرمي)، كما يغطي أيضاً خوارزميات استكشاف البيانات التي عادةً ما يتم اعتبارها صعبة الفهم (على سبيل المثال، خوارزمية نماذج ماركوف المخفية، وخوارزمية القياس المتعدد الأبعاد، وخوارزمية الدعم الآلي المتجه، وخوارزمية تحليل المويجات). كل خوارزميات استكشاف البيانات في هذا الكتاب قد تمّ وصفها بطريقة كاملة وقائمة بذاتها، ومدعّمة بالأمثلة التوضيحية. وبالتالي، فإن هذا الكتاب يتيح للقراء تحقيق نفس المستوى من الفهم الدقيق، وسوف يوفر نفس القدرة من التنفيذ اليدوي بغض النظر عن مستوى صعوبة خوارزميات استكشاف البيانات.

بالنسبة لخوارزميات استكشاف البيانات في كل فصل، يتم سرد قائمة من حزم البرمجيات التي تدعمها. ويتم أيضاً إعطاء بعض التطبيقات لخوارزميات استكشاف البيانات مع المراجع.

المساندة التعليمية:

تتضمن خوارزميات استكشاف البيانات المشمولة في هذا الكتاب مستويات مختلفة من الصعوبة. فالأستاذ الذي يستخدم هذا الكتاب على أنه كتاب تعليمي لمقرر دراسي عن استكشاف البيانات قد يختار الموضوعات المراد تغطيتها بناءً على مستوى المقرر ومستوى صعوبة موضوعات الكتاب. تُعتَبر موضوعات الكتاب في الفصلين ١ و٢ (الأجزاء ٢-١ و٢-٢ فقط)، والفصول ٣، ٤، ٧، ٨، ٩ (الجزء ٩-١ فقط)، والفصول ١٢، ١٦ (الأجزاء من ١٦-١ إلى ١٦-٣ فقط)، والفصل ١٩ (الجزء ١٩-١ فقط)، التي تغطي الأنواع الخمسة من أنماط البيانات، مناسبةً كمقرر خاص بدرجة البكالوريوس، وما تبقى من الموضوعات يُعتبر مناسباً لمقرر في مستوى الدراسات العليا.

وتحتوي نهاية كل فصل على مجموعة من التمارين ذات العلاقة بالموضوعات المطروحة في كل فصل كما يتوافر موقع إلكتروني خاص بالكتاب يحتوي على المواد التعليمية المساندة التالية والتي يمكن الحصول عليها من الناشر:

- دليل حلول التمارين
- العروض التقديمية للمحاضرات، والتي تشمل الخطوط العريضة للموضوعات والأرقام، والجداول، والمعادلات الرياضية

جدير بالذكر أنه يتم استخدام منتج ماتلاب *MATLAB®* لصياغة المعادلات الرياضية في هذا الكتاب. وماتلاب *MATLAB®* هي علامة مسجلة لشركة ماثووركس *MathWorks*. وللحصول على معلومات عن منتج *MATLAB®* يمكن التواصل مع العنوان التالي:

Math Works, Inc.
3 Apple Hill Drive
Natich, MA 1760 - 2098 - USA
Tel: 508 - 647 - 7000
Fax: 508 - 647 - 7001
Email: info@mathworks.com
Web: www.mathworks.com

شكر و تقدير:

أودُ أن أشكر عائلتي، بايجون وأليس، لحبهم و تفهمهم و دعمهم غير المحدود. وأود أن أعرب عن تقديري البالغ لهم لتواجدهم دائماً إلى جانبي وهذا من دواعي سروري حقاً.

وأعرب عن امتناني إلى الدكتور جافريل سالفيندف، الذي كان مُرشدي وصديقي، لتوجيهه لي في مسيرتي الأكاديمية. كما أعرب عن شكري للدكتور غاري هوغ، الذين ساندنى في نواحٍ كثيرة كرئيس للقسم في جامعة ولاية أريزونا.

وأود أيضاً أن أشكر سيندي كاريلى، كبيرة المحررين في دار الطباعة سي آر سي (CRC)، إذ بجهودها وطبيعتها المستجيبة والمساندة و المتفهمة و الداعمة صدر هذا الكتاب ، لقد كان العمل معها فرصة عظيمة. والشكر موصول أيضاً إلى كاري بدفك،كبير منسقي المشاريع في دار الطباعة سي أر سي، وإلى جميع العاملين في الدار الذين ساعدوني في نشر هذا الكتاب.

المؤلفة في سطور:

نونغ يي هي أستاذة في كلية الحاسبات والمعلومات، وهندسة نظم القرار، جامعة ولاية أريزونا، مدينة تيمب ، أريزونا. نونغ يي حاصلة على درجة الدكتوراه في الهندسة الصناعية من جامعة بوردو، لفاييت الغربية بولاية انديانا، و ماجستير في علوم الحاسب الآلي من الأكاديمية الصينية للعلوم، مدينة بكين، جمهورية الصين الشعبية، وعلى درجة البكالوريوس في علوم الحاسب الآلي من جامعة بكين، مدينة بكين، جمهورية الصين الشعبية.

و تشمل إصدارتها كتيب استكشاف البيانات والأنظمة الآمنة للحواسيب والشبكات: النمذجة، والتصميم. وقد نشرت أيضاً أكثر من ٨٠ ورقة عمل في مجلات علمية في مجالات استكشاف البيانات، وتحليل البيانات الإحصائية والنمذجة، وأمن الحاسوب والشبكات، وتحسين جودة الخدمة، ومراقبة الجودة، والتفاعل بين الإنسان والحاسب الآلي، والعوامل البشرية.

الجزء الأول
نظرة عامة على استكشاف البيانات
An overview of Data Mining

١- مقدمة عن البيانات وأنماط البيانات واستكشاف البيانات
Introduction to Data, Data Patterns, and Data Mining

يهدف استكشاف البيانات إلى الكشف عن أنماط البيانات المفيدة من بين كميات هائلة من البيانات. في هذا الفصل، سنوضح بعض الأمثلة لمجموعات من البيانات، واستخدام هذه المجموعات في توضيح أنواع مختلفة من متغيرات البيانات، وأنماط البيانات التي يمكن اكتشافها من البيانات. كما سنتناول في هذا الفصل، ولكن باختصار، خوارزميات استكشاف البيانات حتى نعطي لمحة عن كل نوع من أنماط البيانات. علاوةً على ذلك، سنتناول أيضاً مفهومي البيانات التدريبية والبيانات الاختبارية.

١-١ أمثلة عن مجموعات البيانات الصغيرة
(Examples of Small Data Sets):

لقد مكنت التقنيات الحديثة كأجهزة الحاسوب وأجهزة الاستشعار من أن يتم تسجيل وتخزين وحفظ العديد من الأنشطة مع مرور الزمن، مما نتج عنه تراكم كميات هائلة من البيانات في العديد من المجالات. في هذا الجزء، سنطرح بعض الأمثلة عن مجموعات البيانات الصغيرة التي سيتم استخدامها في هذا الكتاب لشرح مفاهيم استكشاف البيانات والخوارزميات.

ويوضح الجداول ١-١ وحتى الجدول ١-٣ ثلاثة أمثلة لمجموعات بيانات صغيرة تمّ الحصول عليها من مركز (*UCI-Machine Learning Repository*) المتخصص في التعلم الآلي والأنظمة الذكية (*Frank and Asuncion, 2010*). مجموعة بيانات البالون الموضحة في الجدول ١-١ تحتوي على سجلات بيانات لعدد 16 حالة للبالونات. لكل بالون أربع سمات هي: اللون (*Color*)، والحجم (*Size*)، والفعل (*Act*)، والعمر (*Age*). وتحدد سمات البالون هذه ما إذا كان البالون منفوخاً أم لا (*Inflated*). في حين يوضح الجدول ١-٢ مجموعة البيانات الخاصة بتآكل الحلقات الدائرية في مكوك فضاء حيث يحتوي الجدول على سجلات البيانات الخاصة بـ 23 رحلة من رحلات مكوك الفضاء تشالنجر. وهناك أربع سمات لكل رحلة هي: عدد الحلقات الدائرية (*Number of O-Rings*)، درجة حرارة الإطلاق بالفهرنهايت (*Launch Temperature*)، ضغط فحص

التسرب (*Leak-Check Pressure*)، والترتيب الزمني للرحلة (*Temporal Order of Flight*)، والتي يمكن استخدامها لتحديد عدد من الحلقات الدائرية ذات الأحمال الثقيلة (*Number of O-Rings with Stress*). أما مجموعة البيانات الموضحة في الجدول ١-٣، فهي تحتوي على سجلات البيانات لعدد 24 حالة من العدسات لتحديد الملائم منها للمريض. هناك أربع سمات للمريض لكل حالة منها هي: العمر (*Age*)، والتشخيص البصري (*Spectacle Prescription*)، واللابؤرية (*Astigmatic*)، ومعدل خروج الدموع (*Tear Production Rate*)، والتي يمكن استخدامها لتحديد نوع العدسات التي تلائم المريض.

ويوضح الجدول ١-٤ مجموعة البيانات الخاصة باكتشاف الأعطال وتشخيصها في نظام تصنيع معين (*Ye et al., 1993*). يتكون نظام التصنيع من تسع آلات، الآلة الأولى *M1*، الآلة الثانية *M2*،...، الآلة التاسعة *M9*، تقوم بمعالجة وحدات المنتج. ويبين الشكل ١-١ تدفقات عملية الإنتاج التي يتم تنفيذها من خلال الآلات التسع.

الجدول (١-١)

مجموعة بيانات البالون

رقم الحالة Instance	Attribute Variables - متغيرات الخاصية				Target Variable متغير الهدف
	اللون Color	الحجم Size	الفعل Act	العمر Age	خاصية منفوخ Inflated
1	Yellow - أصفر	Small - صغير	Stretch - ممتد	Adult - راشد	T - صحيح
2	Yellow - أصفر	Small - صغير	Stretch - ممتد	Child - طفل	T - صحيح
3	Yellow - أصفر	Small - صغير	Dip - منكمش	Adult - راشد	T - صحيح
4	Yellow - أصفر	Small - صغير	Dip - منكمش	Child - طفل	T - صحيح
5	Yellow - أصفر	Large - كبير	Stretch - ممتد	Adult - راشد	T - صحيح
6	Yellow - أصفر	Large - كبير	Stretch - ممتد	Child - طفل	F - خاطئ
7	Yellow - أصفر	Large - كبير	Dip - منكمش	Adult - راشد	F - خاطئ
8	Yellow - أصفر	Large - كبير	Dip - منكمش	Child - طفل	F - خاطئ
9	Purple - أرجواني	Small - صغير	Stretch - ممتد	Adult - راشد	T - صحيح
10	Purple - أرجواني	Small - صغير	Stretch - ممتد	Child - طفل	F - خاطئ
11	Purple - أرجواني	Small - صغير	Dip - منكمش	Adult - راشد	F - خاطئ
12	Purple - أرجواني	Small - صغير	Dip - منكمش	Child - طفل	F - خاطئ
13	Purple - أرجواني	Large - كبير	Stretch - ممتد	Adult - راشد	T - صحيح
14	Purple - أرجواني	Large - كبير	Stretch - ممتد	Child - طفل	F - خاطئ
15	Purple - أرجواني	Large - كبير	Dip - منكمش	Adult - راشد	F - خاطئ
16	Purple - أرجواني	Large - كبير	Dip - منكمش	Child - طفل	F - خاطئ

هناك بعض وحدات المنتج التي تمر خلال الآلة الأولى $M1$ أولاً، والآلة الخامسة $M5$ ثانياً، والآلة التاسعة $M9$ آخراً، وبعض وحدات المنتج تمر خلال الآلة الأولى $M1$ أولاً، والآلة الخامسة $M5$ ثانياً، والآلة السابعة $M7$ آخراً، وهكذا. هناك تسعة متغيرات، x_i بحيث، $i=1,2,3,4,5,6,7,8,9$، والتي تمثل جودة وحدات المنتج بعد مرورها خلال التسع آلات.

إذا ما اجتازت وحدات المنتج فحص الجودة بعد مرورها بالآلة رقم i، فإن المتغير x_i يأخذ قيمة صفر؛ وخلاف ذلك، فإن x_i يأخذ قيمة واحد. هناك المتغير y، الذي يمثل ما إذا كان النظام به أعطال أم لا. ويكون النظام به أعطال إذا كان أي من التسع آلات بها عطل. إذا لم يكن في النظام أعطال، فإن y تأخذ قيمة صفر؛ وخلاف ذلك، فإن y تأخذ قيمة واحد. هناك تسعة متغيرات، y_i، بحيث، $i=1,2,\ldots,9$، والتي تمثل ما إذا كانت التسع آلات بها أعطال أم لا، على التوالي. إذا لم يكن لدى الآلة i أي عطل، فإن y_i تأخذ قيمة صفر؛ وخلاف ذلك، تأخذ y_i قيمة واحد. وتُستخدم البيانات الخاصة بالكشف عن الأعطال في تحديد ما إذا كان أو لم يكن لدى النظام أعطال استنادا إلى معلومات مستوى الجودة. تستلزم مشكلة الكشف عن الأعطال استخدام متغيرات الجودة التسعة، x_i بحيث، $i=1,2,\ldots,9$، ومتغير أعطال النظام y مشكلة تشخيص الأعطال هي أن تقوم بتحديد الجهاز الذي يحتوي على أعطال بناء على معلومات مستوى الجودة. تستلزم مشكلة تشخيص الأعطال استخدام متغيرات الجودة التسعة، x_i بحيث، $i=1,2,\ldots,9$، ومتغيرات أعطال الجهاز التسعة y_i، بحيث، $i=1,2,\ldots,9$. وقد يكون هناك واحدة أو أكثر من الآلات بها عطل في نفس الوقت، وقد لا تكون هناك أي أعطال بالآلات جميعها. على سبيل المثال، في السجل الأول الذي فيه الآلة الأولى $M1$ بها عطل (فإن y_1 و y تأخذ قيمة واحد، وy_2، وy_3، وy_4، y_5، y_6، y_7، y_8، وy_9 تأخذ قيمة صفر)، ووحدات المنتج بعد المرور على الآلات الأولى $M1$، والخامسة $M5$، والسابعة $M7$، والتاسعة $M9$ قد فشلت في فحص الجودة حيث أخذت متغيرات الجودة x_1، وx_5، وx_7، وx_9 قيمة واحد، ومتغيرات الجودة الأخرى، x_2، وx_3، وx_4، وx_6، وx_8 أخذت قيمة صفر.

الجدول (١-٢)

مجموعة البيانات الخاصة بالحلقات الدائرية في مكوك الفضاء

متغير الهدف Target Variable	Attribute Variables - متغيرات الخاصية				
عدد الحلقات الدائرية ذات الأحمال الثقيلة Number of O-Rings with Stress	الترتيب الزمني للرحلة Temporal Order of Flight	ضغط فحص التسرب Leak-Check Pressure	درجة حرارة الاطلاق Launch Temperature	عدد الحلقات الدائرية Number of O-Rings	رقم الحالة Instance
0	1	50	66	6	1
1	2	50	70	6	2
0	3	50	69	6	3
0	4	50	68	6	4
0	5	50	67	6	5
0	6	50	72	6	6
0	7	100	73	6	7
0	8	100	70	6	8
1	9	200	57	6	9
1	10	200	63	6	10
1	11	200	70	6	11
0	12	200	78	6	12
0	13	200	67	6	13
2	14	200	53	6	14
0	15	200	67	6	15
0	16	200	75	6	16
0	17	200	70	6	17
0	18	200	81	6	18
0	19	200	76	6	19
0	20	200	79	6	20
0	21	200	75	6	21
0	22	200	76	6	22
1	23	200	58	6	23

٢-١ أنواع متغيرات البيانات (Types of Data Variables):

تؤثر أنواع متغيرات البيانات في ماهية خوارزميات استكشاف البيانات التي يمكن تطبيقها على مجموعة معينة من البيانات. هذا الجزء يوضح الأنواع المختلفة لمتغيرات البيانات.

١-٢-١ متغير الخاصية مقابل المتغير الهدف

(Attribute Variable versus Target Variable):

قد يكون لمجموعة بيانات متغيرات خاصية (*Attribute Variables*) ومتغيرات هدف (*Target Variables*)، حيث يتم استخدام قيم متغيرات الخاصية لتحديد قيم متغيرات الهدف. ويمكن أيضاً أن يُطلق على متغيرات الخاصية، ومتغيرات الهدف المتغيرات المستقلة، والمتغيرات التابعة، على التوالي، لتعكس أن قيم المتغيرات الهدف تعتمد على قيم متغيرات الخاصية. في مجموعة البيانات الخاصة بالبالون المذكورة في الجدول ١-١، متغيرات الخاصية هي: اللون (*Color*)، والحجم (*Size*)، والفعل (*Act*)، والعمر (*Age*)، ويوضح المتغير الهدف حالة البالون (منفوخ أو غير منفوخ).

الجدول (١-٣)

مجموعة البيانات الخاصة بالعدسات

رقم الحالة Instance	متغيرات الخاصية - Attributes				متغير الهدف - Target
	العمر Age	التشخيص البصري Spectacle Prescription	اللابؤرية Astigmatic	معدل خروج الدموع Tear Production Rate	العدسات Lenses
1	شاب Young	قُصر النظر Myope	لا No	منخفض Reduced	غير اللاصقة Noncontact
2	شاب Young	قُصر النظر Myope	لا No	طبيعي Normal	اللاصقة الطرية Soft contact
3	شاب Young	قُصر النظر Myope	نعم Yes	منخفض Reduced	غير اللاصقة Noncontact
4	شاب Young	قُصر النظر Myope	نعم Yes	طبيعي Normal	اللاصقة الصلبة Hard contact
5	شاب Young	بُعد النظر Hypermetrope	لا No	منخفض Reduced	غير اللاصقة Noncontact
6	شاب Young	بُعد النظر Hypermetrope	لا No	طبيعي Normal	اللاصقة الطرية Soft contact
7	شاب Young	بُعد النظر Hypermetrope	نعم Yes	منخفض Reduced	غير اللاصقة Noncontact
8	شاب Young	بُعد النظر Hypermetrope	نعم Yes	طبيعي Normal	اللاصقة الصلبة Hard contact
9	ماقبل الشيخوخة Pre-presbyopic	قُصر النظر Myope	لا No	منخفض Reduced	غير اللاصقة Noncontact
10	ماقبل الشيخوخة Pre-presbyopic	قُصر النظر Myope	لا No	طبيعي Normal	اللاصقة الطرية Soft contact
11	ماقبل الشيخوخة Pre-presbyopic	قُصر النظر Myope	نعم Yes	منخفض Reduced	غير اللاصقة Noncontact
12	ماقبل الشيخوخة Pre-presbyopic	قُصر النظر Myope	نعم Yes	طبيعي Normal	اللاصقة الصلبة Hard contact
13	ماقبل الشيخوخة Pre-presbyopic	بُعد النظر Hypermetrope	لا No	منخفض Reduced	غير اللاصقة Noncontact
14	ماقبل الشيخوخة Pre-presbyopic	بُعد النظر Hypermetrope	لا No	طبيعي Normal	اللاصقة الطرية Soft contact

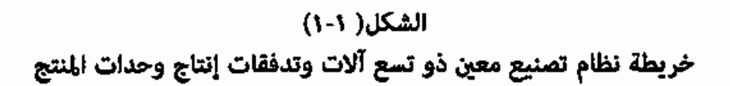
رقم الحالة Instance	العمر Age	التشخيص البصري Spectacle Prescription	اللابؤرية Astigmatic	معدل خروج الدموع Tear Production Rate	العدسات Lenses
	متغيرات الخاصية - Attributes				متغير الهدف - Target
15	ماقبل الشيخوخة Pre-presbyopic	بُعد النظر Hypermetrope	نعم Yes	منخفض Reduced	غير اللاصقة Noncontact
16	ماقبل الشيخوخة Pre-presbyopic	بُعد النظر Hypermetrope	نعم Yes	طبيعي Normal	غير اللاصقة Noncontact
17	الشيخوخة Presbyopic	قُصر النظر Myope	لا No	منخفض Reduced	غير اللاصقة Noncontact
18	الشيخوخة Presbyopic	قُصر النظر Myope	لا No	طبيعي Normal	غير اللاصقة Noncontact
19	الشيخوخة Presbyopic	قُصر النظر Myope	نعم Yes	منخفض Reduced	غير اللاصقة Noncontact
20	الشيخوخة Presbyopic	قُصر النظر Myope	نعم Yes	طبيعي Normal	اللاصقة الصلبة Hard contact
21	الشيخوخة Presbyopic	بُعد النظر Hypermetrope	لا No	منخفض Reduced	غير اللاصقة Noncontact
22	الشيخوخة Presbyopic	بُعد النظر Hypermetrope	لا No	طبيعي Normal	اللاصقة الطرية Soft contact
23	الشيخوخة Presbyopic	بُعد النظر Hypermetrope	نعم Yes	منخفض Reduced	غير اللاصقة Noncontact
24	الشيخوخة Presbyopic	بُعد النظر Hypermetrope	نعم Yes	طبيعي Normal	غير اللاصقة Noncontact

وفي مجموعة البيانات الخاصة بمكوك الفضاء والمذكورة في الجدول ١-٢، فإن متغيرات الخاصية هي: عدد الحلقات الدائرية (*Number of O-rings*)، ودرجة حرارة الإطلاق (*Launch Temperature*)، وضغط فحص التسرب (*Leak-check Pressure*)، والترتيب الزمني للرحلة (*Temporal Order of Flight*)، والمتغير الهدف: هو عدد الحلقات الدائرية ذات الأحمال الثقيلة (*Number of O-rings with stress*).

الجدول (١- ٤)

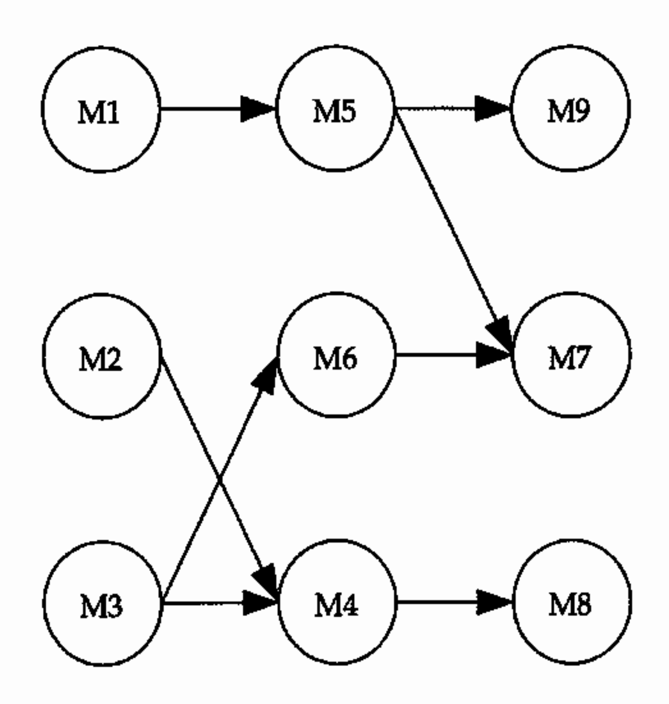
مجموعة البيانات الخاصة باكتشاف الأعطال وتشخيصها في نظام تصنيع معين

Target Variables – متغيرات الهدف										Attribute Variables – متغيرات الخاصية									
Machine Fault – عطل الآلة									عطل النظام (System Fault), y	Quality of Parts – جودة وحدات المنتج									رقم الحالة Instance (الآلة المعطلة Faulty (Machine
y_9	y_8	y_7	y_6	y_5	y_4	y_3	y_2	y_1		x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	
0	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	0	0	1	1(M1)
0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	1	0	2(M2)
0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	3(M3)
0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	4(M4)
0	0	0	0	1	0	0	0	0	1	1	0	1	0	1	0	0	0	0	5(M5)
0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	6(M6)
0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	7(M7)
0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	8(M8)
1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	9(M9)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10(none)



الشكل(١-١)

خريطة نظام تصنيع معين ذو تسع آلات وتدفقات إنتاج وحدات المنتج



قد يكون لبعض مجموعات البيانات متغيرات خاصية فقط. على سبيل المثال، قد تحتوي بيانات العمليات الخاصة بشراء العملاء على العناصر والمواد التي تمّ شراؤها من قبل كل عميل في متجر ما. حيث تمثل العناصر التي تمّ شراؤها متغيرات الخاصية. في كثير من الأحيان تكون الفائدة من بيانات عمليات شراء العملاء هي معرفة العناصر التي يتم شراؤها معاً من قبل العملاء. ويمكن استخدام أنماط اقتران العناصر (أو متغيرات الخاصية) هذه لإعادة تصميم تخطيط المتجر الذي يبيع العناصر وكذلك مساعدة العملاء على التسوق مستقبلاً. إن الاستكشاف والبحث في مثل مجموعة البيانات هذه يستلزم فقط متغيرات الخاصية دون متغيرات الهدف.

١-٢-٢ المتغير النوعي مقابل المتغير الرقمي (Categorical Variable versus Numeric Variable):

يمكن أن يكون للمتغير قيم نوعية أو قيم رقمية. على سبيل المثال، جميع متغيرات الخاصية والمتغير الهدف في مجموعة البيانات الخاصة بالبالون تأخذ قيماً نوعية. فالقيمتان الخاصتان بخاصية اللون هما: الأصفر والأرجواني، تعطيان نوعيتين مختلفتين من اللون. وفي المثال الآخر الخاص ببيانات الحلقات الدائرية لمكوك الفضاء فإن جميع متغيرات الخاصية ومتغيرات الهدف تأخذ قيماً رقمية. على سبيل المثال، قيم متغير الهدف، 0، و1، و2، تمثل عدد الحلقات الدائرية ذات الأحمال. ويمكن استخدام قيم المتغير الرقمي لقياس حجم كمية الاختلافات بين القيم الرقمية. على سبيل المثال، قيمة عدد 2 من الحلقات الدائرية أكبر بمقدار وحدة واحدة من قيمة 1 حلقة دائرية، وأكبر بمقدار وحدتين من قيمة "صفر" حلقة دائرية. وعلى الرغم من ذلك، فإن مقدار كمية الفروقات لا يمكن الحصول عليها من قيم المتغير النوعي. على سبيل المثال، على الرغم من أن اللونين الأصفر والأرجواني يظهران لنا الفرق جلياً بين لونين، فمن غير المناسب تحديد مقياس كمي لذلك الفرق. مثال آخر، الطفل (*Child*) والراشد (*Adult*) هما فئتان نوعيتان مختلفتان خاصة بالعمر. فعلى الرغم من أن كل شخص له / لها عدد من السنوات العمرية، لا يمكننا استخدام الفئتين العمريتين "طفل" و"راشد" للقول بأن "الطفل" أقل عمراً من "الراشد" بمقدار 20، أو 30، أو 40 سنة.

وتنقسم المتغيرات النوعية إلى نوعين فرعيين من المتغيرات: المتغيرات الاسمية (*Nominal Variables*) والمتغيرات الترتيبية (*Ordinal Variables*) (*Tan et al., 2006*). يمكن فرز وترتيب القيم الخاصة بالمتغير الترتيبي، في حين لا يمكن النظر فقط إلى قيم المتغيرات الاسمية على أنها هي ذاتها أو أنها مختلفة. على سبيل المثال، ثلاث قيم للعمر (طفل، راشد، كبير) تجعل هذا المتغير متغيراً ترتيبياً، لأنه يمكن ترتيب القيم (طفل، راشد، كبير) بشكل متصاعد عمرياً. ومع ذلك، لا يمكننا القول بأن فارق العمر بين الطفل والراشد أكبر أو أصغر من فارق العمر بين الراشد والكبير، لأن القيم (طفل، راشد، كبير) هي قيم نوعية وليست قيماً رقمية. وهو ما يعني، أنه على الرغم من أن قيم المتغير الترتيبي يمكن فرزها وترتيبها، فإن هذه القيم نوعية، وفروقها الكمية غير متاحة. اللون هو متغير اسمي حيث إن اللونين الأصفر والأرجواني هما قيمتان مختلفتان، ولكن ترتيب هاتين القيمتين قد يكون غير ذي معنى. يوجد نوعان فرعيان للمتغيرات الرقمية، وهما: متغيرات الفترة (*Interval Variables*)، والمتغيرات النسبية (*Ratio Variables*) (*Tan et al., 2006*).

الفروق الكمية بين قيم متغير الفترة (على سبيل المثال، درجة حرارة الإطلاق F^o) هي ذات معنى، في حين أن كلاً من الفروقات الكمية والنسب بين قيم المتغير النسبي (على سبيل المثال، عدد الحلقات الدائرية ذات الأحمال الثقيلة) هي ذات معنى.

ورسمياً، نرمز لمتغيرات الخاصية بـ $x_1, \ldots, x_p$ ولمتغيرات الهدف، بـ $y_1, \ldots, y_q$. ولتكن $x = (x_1, \ldots x_p)$ و $y = (y_1, \ldots y_q)$. حيث تشير الحالات (أو أمثلة البيانات - *instances*) ومشاهدات البيانات المرصودة (أو الملحوظات المرصودة - *observations*) الخاصة بـ $x_1, \ldots,$ $x_p, y_1, \ldots, y_q$ إلى سجلات البيانات، $(x_1, \ldots, x_p, y_1, \ldots, y_q)$.

١-٣ أنماط البيانات التي يمكن استنباطها من خلال استكشاف البيانات (Data Patterns Learned through Data Mining):

فيما يلي الأنواع الرئيسة لأنماط البيانات التي يتم اكتشافها في مجموعات البيانات باستخدام خوارزميات استكشاف البيانات:

- أنماط التصنيف والتنبؤ
- أنماط الاقتران وأنماط العنقود
- أنماط اختزال البيانات
- الأنماط المتطرفة والشاذة
- الأنماط الزمنية والتسلسلية

وسيتم وصف كل نوع من أنماط البيانات المذكورة أعلاه في الأجزاء التالية.

١-٣-١ أنماط التصنيف والتنبؤ (Classification and Prdiction Patterns):

تُستخدم أنماط التصنيف والتنبؤ في استنباط العلاقات بين متغيرات الخاصية، $(x_1, \ldots, x_p)$ ومتغيرات الهدف $(y_1, \ldots, y_q)$ والمدعومة بمجموعة معطاة من سجلات البيانات، $(x_1, \ldots, x_p, y_1, \ldots, y_q)$ حيث تسمح أنماط التصنيف والتنبؤ بتصنيف أو التنبؤ بقيم المتغيرات الهدف باستخدام قيم متغيرات الخاصية.

على سبيل المثال، جميع سجلات البيانات الـ 16 في مجموعة البيانات الخاصة بالبالون والمذكورة في الجدول ١-١ تدعم العلاقة التالية لمتغيرات الخاصية، اللون (*Color*)، والحجم (*Size*)، والفعل (*Act*)، والعمر (*Age*) مع متغير الهدف "منفوخ" (*Inflated*) (حيث تشير القيمة "*T*" إلى "*True*" أي "صحيح": أي أنّ البالون منفوخ و تشير القيمة "*F*" إلى "*False*" أي "خاطئ": أي أنّ البالون غير منفوخ):

IF (Color = Yellow AND Size = Small) OR (Age = Adult AND Act = Stretch), THEN Inflated = T; OTHERWISE, Inflated = f.
إذا كان (اللون = أصفر، و الحجم = صغير) أو (العمر = راشد و الفعل = ممتد)، **إذن** تكون خاصية منفوخ = T (أي "صحيح")؛ **وإلا** تكون خاصية منفوخ = F (أي "خاطئ").

العلاقة المذكورة أعلاه تسمح لنا بتصنيف بالون ما إلى قيمة نوعية لمتغير الهدف باستخدام قيمة محددة لمتغيرات الخاصية: اللون (*Color*)، والحجم (*Size*)، والفعل (*Act*)، والعمر (*Age*). وبالتالي، فإن هذه العلاقة تعطينا نمط بيانات تسمح لنا بإجراء التصنيف للبالون. وعلى الرغم من أنه يمكننا استخلاص نمط العلاقة هذا عن طريق فحص سجلات البيانات الـ 16 في مجموعة بيانات البالون، إلا أن استخلاص هذا النمط يدوياً من مجموعة كبيرة جداً من البيانات المختلطة ببيانات مشوشة قد يكون مهمةً صعبة. إن استخدام خوارزمية استكشاف البيانات يمكننا من التعلم من مجموعة كبيرة من البيانات بشكل تلقائي.

وبمثال آخر، فإن النموذج الخطي التالي يلائم 23 سجلاً بيانياً لمتغير الخاصية، وهو درجة حرارة الإطلاق (*Launch Temperature*)، والمتغير الهدف: عدد الحلقات الدائرية ذات الأحمال الثقيلة (*Number of O-rings with stress*)، في مجموعة البيانات الخاصة بالحلقات الدائرية في مكوك الفضاء المذكورة في الجدول ١-٢:

$$y = -0.05746\,x + 4.301587 \qquad (1\text{-}1)$$

حيث:

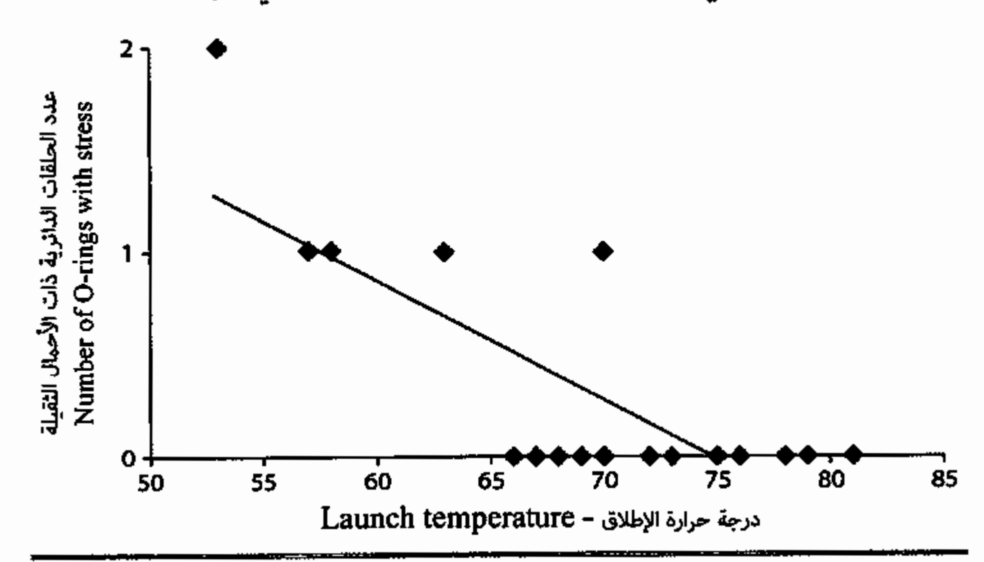
y تشير إلى المتغير الهدف: عدد الحلقات الدائرية ذات الأحمال الثقيلة (*Number of O-rings with stress*)

x تشير إلى متغير الخاصية، وهو درجة حرارة الإطلاق (*Launch Temperature*)

يوضح الشكل ١-٢ قيم درجة حرارة الإطلاق، وعدد الحلقات الدائرية ذات الأحمال الثقيلة في الـ 23 سجلاً بيانياً، ويوضح الخط الملائم الموضح في المعادلة الخطية ١-١. ويبين الجدول ١-٥ قيمة الخاصية: الحلقات الدائرية ذات الأحمال الثقيلة، لكل سجل من سجلات البيانات التي تمّ التنبؤ بها من قيمة درجة حرارة الإطلاق باستخدام نموذج العلاقة الخطية لدرجة حرارة الإطلاق مع عدد الحلقات الدائرية ذات الأحمال الثقيلة في المعادلة ١-١. باستثناء اثنين من سجلات البيانات للحالتين 2 و11، فإن النموذج الخطي في المعادلة ١-١ يجسد العلاقة بين درجة حرارة الإطلاق مع عدد الحلقات الدائرية ذات الأحمال الثقيلة بشكل جيد، إذ إنه كلما انخفضت قيمة درجة حرارة الاطلاق زادت قيمة الحلقات الدائرية ذات الأحمال الثقيلة. ويتضح أن القيمة المتوقعة الأعلى لعدد الحلقات الدائرية ذات الأحمال الثقيلة تظهر جلياً في سجل البيانات رقم 14 مع 2 من الحلقات الدائرية بها أحمال حرارية.

الشكل(١-٢)

النموذج الملائم للعلاقة الخطية الخاصة بدرجة حرارة الإطلاق مع عدد الحلقات الدائرية ذات الأحمال الثقيلة في مجموعة البيانات الخاصة بالحلقات الدائرية في مكوك الفضاء



القيمتان اللتان تمّ التنبؤ بهما في النطاق المتوسط، 1.026367 و 0.681607 ، تظهران بوضوح في اثنين من سجلات البيانات أرقام 9، 10 في الجدول ١-٥ مع واحد من الحلقات الدائرية ذات الأحمال الثقيلة. وتظهر القيم المتوقعة في نطاق منخفض من -0.352673 إلى 0.509227 لجميع سجلات البيانات التي يبلغ عدد الحلقات الدائرية ذات الأحمال الثقيلة بها صفراً. كما يكشف المعامل السلبي لـ x -0.05746، في المعادلة ١-١ هذه العلاقة. وبالتالي، فإن العلاقة الخطية في المعادلة ١-١ تعطي نمطاً للبيانات يتيح لنا التنبؤ بالمتغير الهدف (عدد الحلقات الدائرية ذات الأحمال الثقيلة)، من متغير الخاصية (درجة حرارة الإطلاق) في مجموعة البيانات الخاصة بالحلقات الدائرية في مكوك الفضاء.

يمكن تمثيل أنماط التصنيف والتنبؤ، التي تصور علاقة متغيرات الخاصية، $(x_1, \ldots, x_p)$ مع متغيرات الهدف، $(y_1, \ldots, y_q)$، بالشكل العام $y=F(x)$. بالنسبة لمجموعة بيانات البالون، فإن أنماط التصنيف (*Classification Patterns*) الخاصة بـF تأخذ شكل قواعد القرار. وبالنسبة لمجموعة البيانات الخاصة بعدد الحلقات الدائرية في مكوك الفضاء، فإن أنماط التنبؤ (*Prediction Patterns*) لـ F تأخذ شكل النموذج الخطي. وبشكل عام، يُستخدم مصطلح "أنماط التصنيف" إذا كان المتغير الهدف هو متغير نوعي، أما مصطلح "أنماط التنبؤ" فيَستخدم إذا كان المتغير الهدف هو متغير رقمي.

الجدول(١-٥)

القيمة المتوقعة لعدد الحلقات الدائرية ذات الأحمال الثقيلة

رقم الحالة Instance	متغير الخاصية Attribute Variable	Target Variable - متغير الهدف	
	درجة حرارة الإطلاق Launch Temperature	عدد الحلقات الدائرية ذات الأحمال الثقيلة Number of O-Rings with Stress	القيمة المتوقعة بها لعدد الحلقات الدائرية ذات الأحمال الثقيلة Predicted Value of O-Rings with Stress
1	66	0	0.509227
2	70	1	0.279387
3	69	0	0.336847
4	68	0	0.394307
5	67	0	0.451767
6	72	0	0.164467
7	73	0	0.107007
8	70	0	0.279387
9	57	1	1.026367
10	63	1	0.681607
11	70	1	0.279387
12	78	0	−0.180293
13	67	0	0.451767
14	53	2	1.256207
15	67	0	0.451767
16	75	0	−0.007913
17	70	0	0.279387
18	81	0	−0.352673
19	76	0	−0.065373
20	79	0	−0.237753
21	75	0	−0.007913
22	76	0	−0.065373
23	58	1	0.968907

يستعرض الجزء الثاني من الكتاب خوارزميات استكشاف البيانات التالية التي يتم استخدامها لاستنباط أنماط التصنيف والتنبؤ من البيانات:

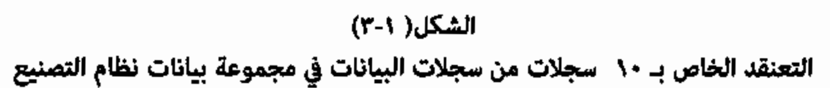
- نماذج الانحدار في الفصل ٢
- مصنْف بييز البسيط في الفصل ٣
- أشجار القرار والانحدار في الفصل ٤
- الشبكات العصبية الصناعية للتصنيف والتنبؤ في الفصل ٥
- الدعم الآلي المتجه في الفصل ٦
- مصنّف أقرب k - مجاور والتعنقد المراقب في الفصل ٧

توضح الفصول ٢٠، و٢١، و٢٣، الموجودة في كتيب استكشاف البيانات (*Ye, 2003*) (*The Handbook of Data Mining*) والفصلان ١٢ و١٣ في كتاب الأنظمة الآمنة للحواسيب والشبكات: النمذجة والتحليل والتصميم (*Ye, 2008*) (*Secure Computer and Network Systems: Modeling, Analysis and Design*) التطبيقات الخاصة بخوارزميات التصنيف والتنبؤ لبيانات الأداء الإنساني، والبيانات النصية، والبيانات العلمية والهندسية، والبيانات الخاصة بالحاسوب والشبكات.

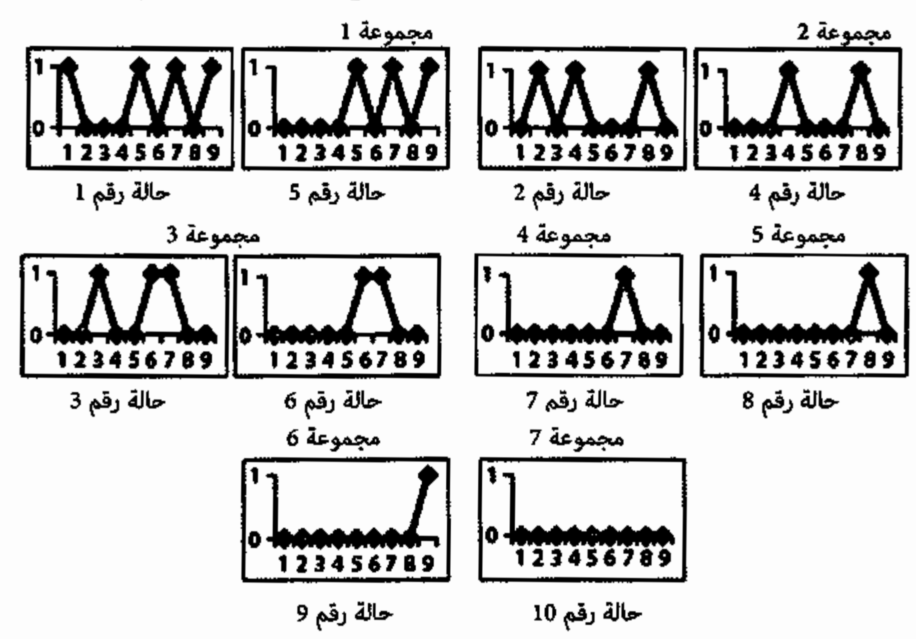
١-٣-٢ أنماط الاقتران وأنماط العنقود (Cluster and Association Patterns):

عادةً ما تستلزم أنماط الاقتران وأنماط العنقود متغيرات الخاصية فقط، ($x_1, \ldots, x_p$)، (يُطلق مصطلح العنقود -*cluster*- ليَشير إلى المجموعة المتشابهة من سجلات البيانات). وتحتوي أنماط العنقود على مجموعات من سجلات البيانات المتماثلة بحيث تكون سجلات البيانات في مجموعة واحدة متشابهة، ولكن هناك اختلافات أكبر عن سجلات البيانات في مجموعة أخرى. وبعبارة أخرى، فإن أنماط العنقود تكشف عن أنماط التشابه والاختلاف بين سجلات البيانات. أما أنماط الاقتران فيتم تشكيلها على أساس التلازم والتزامن في حدوث العناصر الموجودة في سجلات البيانات، (يُطلق مصطلح الاقتران -*association*- ليَشير إلى ارتباط وقوع أو حدوث العناصر أو المتغيرات الموجودة في سجلات البيانات). في بعض

الأحيان، تُستخدَم أيضاً المتغيرات الهدف، $(y_1, ..., y_q)$ في التعنقد، ولكن يتم التعامل معها بنفس الطريقة التي يتم التعامل بها مع متغيرات الخاصية.


الشكل(١-٣)

التعنقد الخاص بـ ١٠ سجلات من سجلات البيانات في مجموعة بيانات نظام التصنيع



على سبيل المثال، يمكن تجميع ١٠ من سجلات البيانات الموجودة في مجموعة بيانات نظام التصنيع والموضحة في الجدول ١-٤ في سبع مجموعات، كما هو مبين في الشكل ١-٣. حيث يوضح المحور الأفقي لكل رسم بياني في الشكل ١-٣ متغيرات الجودة التسعة، ويوضح المحور الرأسي قيمة متغيرات الجودة التسعة تلك. هناك ثلاث مجموعات تتكون من أكثر من سجل واحد من سجلات البيانات: المجموعة الأولى (*Group 1*)، والمجموعة الثانية (*Group 2*)، والمجموعة الثالثة (*Group 3*). ضمن كل مجموعة من هذه المجموعات، تبدو سجلات البيانات متشابهة مع اختلاف القيم في واحدة فقط من متغيرات الجودة التسعة. إن إضافة أي سجل بيانات آخر إلى كل مجموعة من هذه المجموعات الثلاث يجعل

المجموعة لديها على الأقل اثنين من سجلات البيانات بها قيم مختلفة في أكثر من متغير جودة واحد.

لنفس مجموعة بيانات نظام التصنيع، فإن متغيرات الجودة، x_4 وx_8 مقترنة ببعضها بشكل عالٍ لأن لديها نفس القيمة في جميع سجلات البيانات باستثناء السجل رقم 8. وهناك أزواج أخرى من المتغيرات، على سبيل المثال، x_5 وx_9 والتي ترتبط ببعضها إلى حد كبير لنفس السبب. هذه هي بعض أنماط الاقتران الموجودة في مجموعة بيانات نظام التصنيع في الجدول ١-٤.

كما يناقش الجزء الثالث من الكتاب خوارزميات استكشاف البيانات التالية التي يتم استخدامها في استنباط أنماط العنقود وأنماط الاقتران من البيانات:

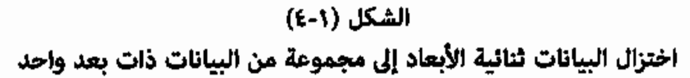
- التعنقُد الهرمي في الفصل (٨).
- التعنقد حول K من المتوسطات والتعنقد على أساس الكثافة في الفصل (٩).
- خريطة التنظيم الذاتي في الفصل (١٠).
- التوزيعات الاحتمالية للبيانات أحادية المتغير في الفصل (١١).
- قواعد الاقتران في الفصل (١٢).
- شبكات بييز في الفصل (١٣).

وتتناول الفصول ١٠، و٢١، و٢٢، و٢٧، الموجودة في كتيب استكشاف البيانات (*Ye, 2003*)، التطبيقات الخاصة بخوارزميات العناقيد لبيانات سلّة السوق، وبيانات الدخول إلى شبكة الإنترنت، والبيانات النصية، والبيانات الجغرافية المكانية، وبيانات الصور. بينما يتناول الفصل ٢٤، الموجود في كتيب استكشاف البيانات (*Ye, 2003*)، التطبيق الخاص بخوارزمية قاعدة الاقتران لبيانات تركيب البروتين.

١-٣-٣ أنماط اختزال البيانات (Data Reduction Patterns):

تبحث أنماط اختزال البيانات عن عدد قليل من المتغيرات التي يمكن استخدامها لتمثيل مجموعة من البيانات ذات عدد أكبر بكثير من المتغيرات. حيث إن المتغير الواحد يعطي بعداً واحداً من البيانات، وتسمح أنماط اختزال البيانات لمجموعة من البيانات ذات أبعاد

كثيرة أن يتم تمثيلها في مجموعة بيانات ذات أبعاد أقل. على سبيل المثال، يوضح الشكل ١-٤، عشرة سجلات بيانات في فضاء ثنائي الأبعاد (x, y)، حيث $y=x_2 , x=1,2,...,10$. يمكن تمثيل مجموعة البيانات الثنائية الأبعاد هذه كمجموعة بيانات ذات بعد واحد بحيث تكون z محوراً، وتكون z مرتبطة بالمتغيرات الأصلية، y وx على النحو التالي:

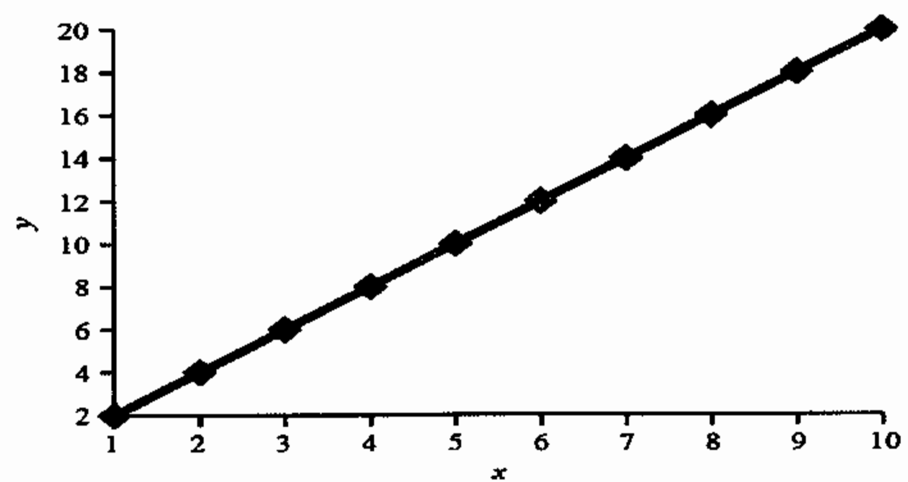


$$z = x * \sqrt{1^2 + 1 * \left(\frac{y}{x}\right)^2}. \qquad (٢-١)$$

وتكون نقاط البيانات العشر لـ z هي: 2.236، 4.472، 6.708، 8.944، 11.180، 13.416، 15.652، 17.889، 20.125، و22.361.

الشكل (١-٤)

اختزال البيانات ثنائية الأبعاد إلى مجموعة من البيانات ذات بعد واحد



أما الجزء الرابع من الكتاب، فيستعرض خوارزميات استكشاف البيانات التالية التي يتم استخدامها لاكتشاف أنماط اختزال البيانات من البيانات:

- تحليل المكونات الرئيسية (الفصل ١٤).
- القياس المتعدد الأبعاد (الفصل ١٥).

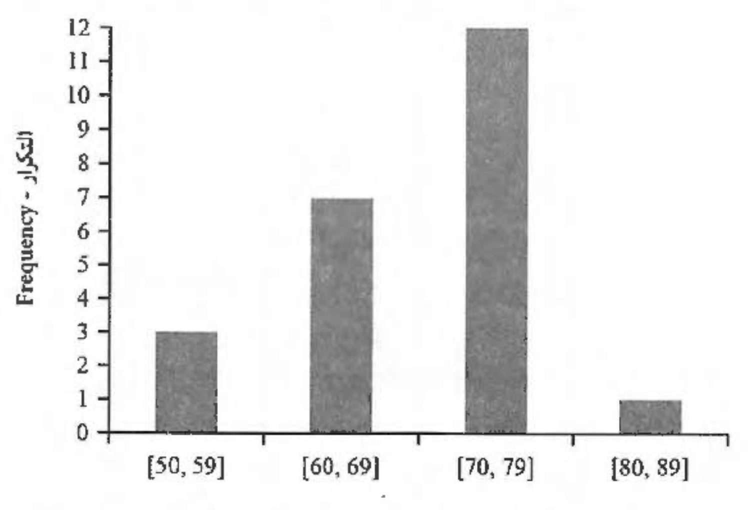
ويتناول الفصلان ٢٣ و٨، الموجود في كتيب استكشاف البيانات (*Ye, 2003*)، تطبيقات تحليل المكون الرئيسي لبيانات البراكين وبيانات العلوم والهندسة.

١-٣-٤ الأنماط المتطرفة والشاذة (Outlier and Anomaly Patterns):

القيم المتطرفة (outliers) والشاذة (anomaly) هي نقاط البيانات التي تختلف إلى حد كبير عن المعيار العام للبيانات. ويمكن تعريف المعيار العام للبيانات بعدة طرق. على سبيل المثال، يمكن تعريف المعيار على أنه نطاق القيم الذي تشغله غالبية نقاط البيانات، ونقطة البيانات ذات القيمة التي تكون خارج هذا النطاق، يمكن اعتبارها قيمةً متطرفة. يوضح الشكل ١-٥ رسماً بيانياً لتكرار قيم درجة حرارة الإطلاق الخاص بنقط البيانات في مجموعة بيانات مكوك الفضاء المذكورة في الجدول ١-٢. هناك ثلاث قيم من قيم درجة حرارة الإطلاق في النطاق *[50, 59]*، وعدد سبع قيم في النطاق *[60, 69]*، وعدد اثنتي عشرة قيمة في النطاق *[70, 79]*، وقيمة واحدة فقط في النطاق *[80, 89]*. وبالتالي، فإن غالبية قيم درجة حرارة الإطلاق هي في النطاق *[50, 79]*. ويمكن اعتبار القيمة 81 في السجل 18 قيمةً متطرفةً أو شاذةً.

الشكل (١-٥)

الرسم البياني التكراري لدرجات حرارة الإطلاق في مجموعة بيانات مكوك الفضاء



ويستعرض الجزء الخامس من الكتاب خوارزميات استكشاف البيانات التالية التي تُستخدَم لتحديد بعض المعايير الإحصائية للبيانات، وللكشف عن القيم المتطرفة والشاذة وفقاً لتلك المعايير الإحصائية:

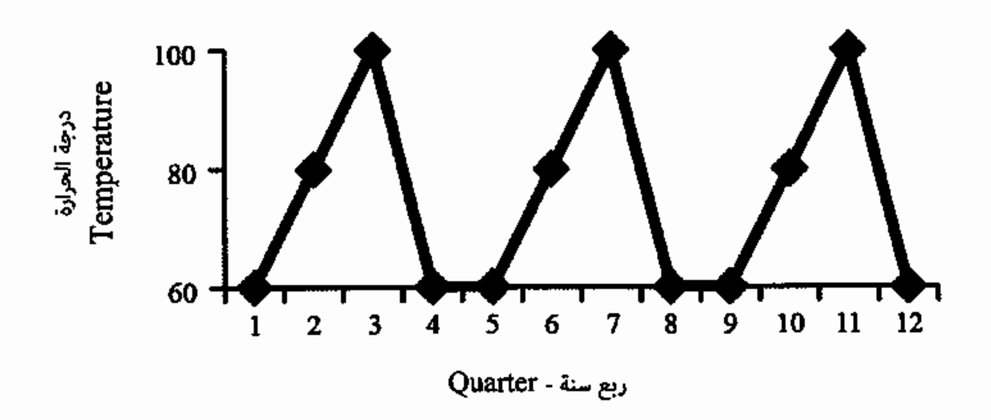
- مخطط التحكم أحادي المتغير في الفصل ١٦
- مخطط التحكم متعدد المتغيرات في الفصل ١٧

تقدم الفصول ٢٦ و٢٨، الموجودة في كتيب استكشاف البيانات (*Ye, 2003*)، والفصل ١٤ الذي يدور حول الأنظمة الآمنة للحواسيب والشبكات: النمذجة والتحليل والتصميم (*Ye, 2008*)، التطبيقات الخاصة بخوارزميات الكشف عن البيانات المتطرفة والشاذة في بيانات القطاع الصناعي وبيانات الحواسيب والشبكات.

١-٣-٥ الأنماط الزمنية والتسلسلية (Sequential and Temporal Patterns):

تكشف الأنماط الزمنية والتسلسلية عن الأنماط الموجودة في سلسلة نقاط أو سجلات البيانات. إذا تمّ تعريف التسلسل على أنه الوقت الذي جُمعت خلاله نقاط البيانات، فإننا نطلق على سلسلة نقاط البيانات "سلسلة الزمن". يوضح الشكل ١-٦ السلسلة الزمنية لقيم درجات الحرارة في مدينة ما كل ثلاثة شهور لمدة ثلاث سنوات.

الشكل(١-٦)
درجة حرارة الطقس كل ثلاثة شهور لمدة ٣ سنوات



الجدول (١-٦)

مجموعة بيانات اختبارية لنظام تصنيع معين لاكتشاف وتشخيص الأعطال

Target Variables – متغيرات الهدف										Attribute Variables – متغيرات الخاصية									
Machine Fault – عطل الآلة									عطل النظام (System Fault), y	Quality of Parts – جودة وحدات المنتج									رقم الحالة Instance (الآلة المعطلة)
y_9	y_8	y_7	y_6	y_5	y_4	y_3	y_2	y_1		x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	(Faulty Machine
0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	0	1	1	1 (M1,M2)
0	0	0	0	0	0	1	1	0	1	0	1	1	1	0	1	1	1	0	2 (M2,M3)
0	0	0	0	0	0	1	0	1	1	1	0	1	1	1	0	1	0	1	3 (M1,M3)
0	0	0	0	0	1	0	0	1	1	1	1	1	0	1	1	0	0	1	4 (M1,M4)
0	0	0	1	0	0	0	0	1	1	1	0	1	1	1	0	0	0	1	5 (M1,M6)
0	0	0	1	0	0	0	1	0	1	0	1	1	1	0	1	0	1	0	6 (M2,M6)
0	0	0	0	1	0	0	1	0	1	0	1	1	0	1	1	0	1	0	7 (M2,M5)
0	0	0	0	1	0	1	0	0	1	1	0	1	1	1	0	1	0	0	8 (M3,M5)
0	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	9 (M4,M7)
0	1	0	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	10 (M5,M8)
1	0	0	0	0	0	1	0	0	1	1	1	1	1	0	1	1	0	0	11 (M3,M9)
0	1	0	0	0	0	0	0	1	1	1	1	1	0	1	0	0	0	1	12 (M1,M8)
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	13 (M1,M2,M3)
0	0	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	0	14 (M2,M3,M5)
1	0	0	0	0	0	1	1	0	1	1	1	1	1	0	1	1	1	0	15 (M2,M3,M9)
1	0	0	1	0	0	0	0	1	1	1	1	1	1	1	0	0	0	1	16 (M1,M6,M8)

هناك نمط دوري لدرجات الحرارة: ٦٠، ٨٠، ١٠٠، و ٦٠، والذي يتكرر كل عام. يمكن اكتشاف مجموعة متنوعة من الأنماط الزمنية والتسلسلية باستخدام خوارزميات استكشاف البيانات في الجزء السادس من الكتاب، بما في ذلك:

- تحليل الارتباط الذاتي وسلاسل الزمن في الفصل (١٨).
- نماذج سلسلة ماركوف ونماذج ماركوف المخفية في الفصل (١٩).
- تحليل المويجات في الفصل (٢٠).

و تتناول الفصول ١٠، و١١، و١٦، الموجودة في كتاب الأنظمة الآمنة للحواسيب والشبكات: النمذجة والتحليل والتصميم (*Ye, 2008*)، التطبيقات الخاصة بخوارزميات استكشاف نمط تسلسلي وزمني لبيانات الحاسب والشبكات، لكشف الهجمات الحاسوبية عبر الإنترنت.

١-٤ البيانات التدريبية والبيانات الاختبارية (Training Data and Test Data):

مجموعة البيانات التدريبية (أو الاستكشافية) هي مجموعة من سجلات البيانات التي يتم استخدامها لمعرفة واكتشاف أنماط البيانات. بعد اكتشاف أنماط البيانات، ينبغي اختبارها لمعرفة إمكانية تعميمها على مجموعة واسعة من سجلات البيانات، بما في ذلك تلك التي تختلف عن سجلات البيانات التدريبية. وتُستخدم مجموعة البيانات الاختبارية لهذا الغرض، بالإضافة إلى احتوائها على سجلات بيانات جديدة ومختلفة. على سبيل المثال، يبين الجدول ١-٦ مجموعة بيانات اختبارية لتصنيع نظام معين واكتشاف أعطاله وتشخيصها. وتحتوي مجموعة البيانات التدريبية لنظام التصنيع هذا والمذكورة في الجدول ١-٤ على سجلات بيانات خاصة بتسع أعطال أحادية الآلة، وحالة واحدة لآلة بدون أعطال. تحتوي مجموعة البيانات الاختبارية في الجدول ١-٦ على سجلات بيانات لبعض الأعطال ثنائية الآلة وثلاثية الآلة أيضاً.

التمارين (Exercises):

١-١ أوجد وقم بوصف مجموعة بيانات تحتوي على ٢٠ سجل بيانات على الأقل، والتي سبق استخدامها في تطبيق لاستكشاف البيانات لغرض اكتشاف أنماط التصنيف، على أن تحتوي مجموعة البيانات هذه على العديد من متغيرات الخاصية النوعية، ومتغير هدف نوعي.

١-٢ أوجد وقم بوصف مجموعة بيانات تحتوي على ٢٠ سجل بيانات على الأقل، والتي سبق استخدامها في تطبيق لاستكشاف البيانات لغرض اكتشاف أنماط التنبؤ، على أن تحتوي مجموعة البيانات هذه على العديد من متغيرات الخاصية الرقمية، ومتغير هدف رقمي.

١-٣ أوجد وقم بوصف مجموعة بيانات تحتوي على ٢٠ سجل بيانات على الأقل، والتي سبق استخدامها في تطبيق لاستكشاف البيانات لغرض اكتشاف أنماط العنقود، على أن تحتوي مجموعة البيانات هذه على متغيرات الخاصية متعددة ورقمية.

١-٤ أوجد وقم بوصف مجموعة بيانات تحتوي على ٢٠ سجل بيانات على الأقل، والتي سبق استخدامها في تطبيق لاستكشاف البيانات لغرض اكتشاف أنماط الاقتران على أن تحتوي مجموعة البيانات هذه على عدة متغيرات نوعية.

١-٥ أوجد وقم بوصف مجموعة بيانات تحتوي على ٢٠ سجل بيانات على الأقل، والتي سبق استخدامها في تطبيق لاستكشاف البيانات لغرض اكتشاف أنماط اختزال البيانات، وحدد نوع (أنواع) متغيرات البيانات في مجموعة البيانات هذه.

١-٦ أوجد وقم بوصف مجموعة بيانات تحتوي على ٢٠ سجل بيانات على الأقل، والتي سبق استخدامها في تطبيق لاستكشاف البيانات لغرض اكتشاف الأنماط المتطرفة والشاذة، وحدد نوع (أنواع) متغيرات البيانات في مجموعة البيانات هذه.

١-٧ أوجد وقم بوصف مجموعة بيانات تحتوي على ٢٠ سجل بيانات على الأقل، والتي سبق استخدامها في تطبيق لاستكشاف البيانات لغرض اكتشاف الأنماط الزمنية والتسلسلية، وحدد نوع (أنواع) متغيرات البيانات في مجموعة البيانات هذه.

الجزء الثاني
خوارزميات لاستكشاف أنماط التصنيف والتنبؤ

Algorithms for Mining Classification and Prediction Patterns

٢- نماذج الانحدار الخطية وغير الخطية
Linear and Nonlinear Regression Models

تعمل نماذج الانحدار على توضيح الكيفية التي يتغير بها واحدٌ أو أكثر من متغيرات الهدف تبعاً لتغير واحد أو أكثر من متغيرات الخاصية. ويمكن استخدامها للتنبؤ بقيم متغيرات الهدف باستخدام قيم متغيرات الخاصية. وفي هذا الفصل، سنتناول نماذج الانحدار الخطية وغير الخطية. كما سنناقش في هذا الفصل طريقة المربعات الصغرى (*least squares method*) وطريقة الإمكان الأكبر (*maximum likelihood method*) لتقدير المعلمات في نماذج الانحدار. بالإضافة إلى ذلك، سيتم تقديم قائمة من الحزم البرمجية التي تدعم بناء نماذج الانحدار.

٢-١ نماذج الانحدار الخطي (Linear Regression Models):

يحتوي نموذج الانحدار الخطي البسيط، على متغير هدف واحد y فقط ومتغير خاصية واحد x فقط كما هو موضح أدناه:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (٢-١)$$

حيث إنّ:

(x_i, y_i) تشير إلى الملحوظة المرصودة رقم i لكل من x و y.

ε_i يمثل الخطأ العشوائي (على سبيل المثال، خطأ القياس) الذي يسهم في الملحوظة المرصودة رقم i الخاصة بالمتغير y.

بالنسبة لقيمة معينة لـ x_i فإنّ كلاً من y_i و ε_i يعد متغيرات عشوائية يمكن أن يتبع قيمها توزيعاً احتمالياً كما هو موضح في الشكل ٢-١. وبعبارة أخرى، لنفس قيمة x يمكن ملاحظة قيم مختلفة لـ y و ε في أوقات مختلفة. يوجد ثلاثة افتراضات خاصة بـ ε_i:

١- $E(\varepsilon_i) = 0$ وهو ما يعني أن متوسط الخطأ العشوائي ε_i يساوي الصفر.
٢- $var(\varepsilon_i)=\sigma^2$ وهو ما يعني أن الأخطاء العشوائية لها تباين ثابت يساوي σ^2.

٣- $cov(\varepsilon_i , \varepsilon_j)=0$ حيث $i \neq j$ وهو ما يعني أن التباين المصاحب (*covariance*) لكل من $(\varepsilon_i , \varepsilon_j)$ لأي ملحوظتين مرصودتين بيانيتين مختلفتين (الملحوظة رقم i والملحوظة رقم j) يساوي صفراً.

هذه الافتراضات تعني أن:

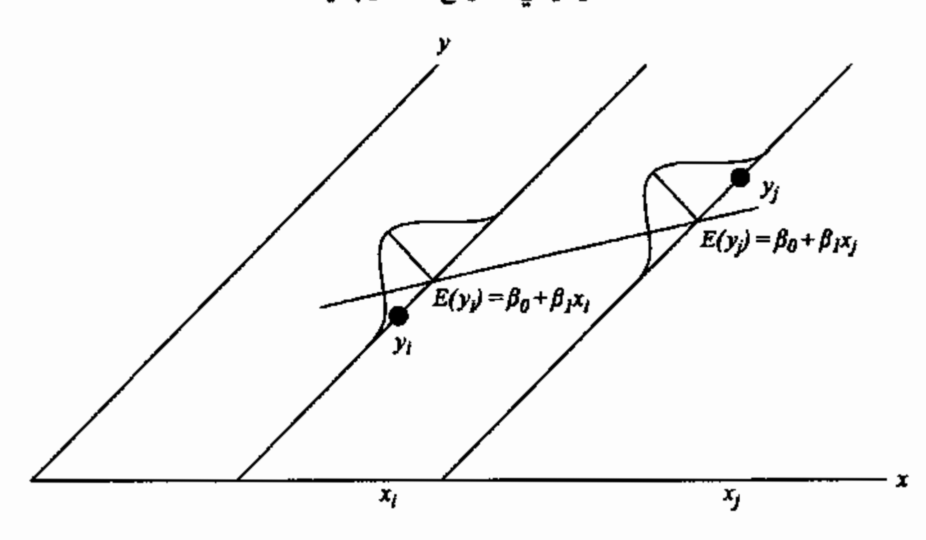
١- $E(y_i) = \beta_0 + \beta_1 x_i$

٢- $var(y_i) = \sigma^2$

٣- $cov(y_i , y_j) = 0$ لأي ملحوظتين مرصودتين بيانيتين مختلفتين لـ y، الملحوظة رقم i والملحوظة رقم j.

الشكل (٢-١)

مثال توضيحي لنموذج انحدار بسيط



ويمكن توسيع نموذج الانحدار الخطي البسيط في المعادلة ٢-١ ليشمل متغيرات خاصية متعددة:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i \qquad (٢-٢)$$

حيث إن:

p هو عدد صحيح أكبر من 1.

$x_{i,j}$ تشير إلى الملحوظة المرصودة رقم i لمتغير الخاصية رقم j.

نماذج الانحدار الخطي في المعادلتين ٢-١ و٢-٢ هي خطية بالمعلمات: $\beta_0, \ldots, \beta_p$ ، ومتغيرات الخاصية: $x_{i,1}, \ldots, x_{i,p}$ وبشكل عام، نماذج الانحدار الخطي هي خطية في المعلمات ولكنها ليست بالضرورة خطية في متغيرات الخاصية. نموذج الانحدار التالي متعدد الحدود للمتغير x_1 هو أيضاً نموذج انحدار خطي:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,1}^k + \varepsilon_i \qquad (٣-٢)$$

حيث إن k هو عدد صحيح أكبر من 1. ويأتي الشكل العام لنموذج الانحدار الخطي كما يلي:

$$y_i = \beta_0 + \beta_1 \Phi_1(x_{i,1}, \ldots, x_{i,p}) + \ldots + \beta_k \Phi_k(x_{i,1}, \ldots, x_{i,p}) + \varepsilon_i \qquad (٤-٢)$$

حيث إن Φ_l ، $l=1,..,k$ هي دالة خطية أو غير خطية تستلزم واحداً أو أكثر من المتغيرات $x_i, \ldots, x_p$ وفيما يلي مثال آخر لنموذج انحدار خطي بمعلماته:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 \log x_{i,1} x_{i,2} + \varepsilon_i \qquad (٥-٢)$$

٢-٢ طريقة المربعات الصغرى وطريقة الإمكان الأكبر لتقدير المعلمة
(Least-Squares Method and Maximum Likelihood Method of Parameter Estimation):

حتى يتم ملاءمة نموذج انحدار خطي مع مجموعة من البيانات التدريبية أو الاستكشافية (x_i, y_i)، حيث $x_i=(x_{i,1}, \dots, x_{i,p})$، $i=1, \dots, n$، فإننا نحتاج إلى تقدير المعلمات βs (المَعلَمات: مفردها مَعلَمة وهي عبارة عن عامل متغير قابل للقياس في نظام معادلات معين). عادةً ما يتم استخدام طريقة المربعات الصغرى وطريقة الإمكان الأكبر لتقدير المعلمات βs. وسوف يتم توضيح كلتا الطريقتين باستخدام نموذج الانحدار الخطي البسيط في المعادلة ٢-١.

تبحث طريقة المربعات الصغرى عن قيم للمعلمات β_0 وβ_1 التي تقلل من مجموع الأخطاء التربيعية (SSE) بين القيم المستهدفة الفعلية $(y_i, i=1,..,n)$ والقيم المستهدفة المقدَّرة $(\hat{y}_i, i=1,\dots,n)$ باستخدام المعلمات المقدَّرة $\hat{\beta}_0$ و $\hat{\beta}_1$. مجموع الأخطاء التربيعية (SSE) عبارة عن دالة لكل من $\hat{\beta}_0$ و $\hat{\beta}_1$:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \qquad (٢\text{-}٦)$$

يجب أن تكون قيمة الاشتقاق الجزئي لـ SSE فيما يتعلق بـ $\hat{\beta}_0$ و $\hat{\beta}_1$ صفراً عند النقطة التي يتم فيها تصغير SSE. ومن ثم، فإن قيم $\hat{\beta}_0$ و $\hat{\beta}_1$ التي تُصغِّر قيمة SSE يتم الحصول عليها باشتقاق SSE بالنسبة لـ $\hat{\beta}_0$ و $\hat{\beta}_1$، ووضع هذه الاشتقاقات الجزئية مساويةً للصفر:

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \qquad (٢\text{-}٧)$$

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \qquad (٢\text{-}٨)$$

يتم تبسيط المعادلات ٢-٧ و٢-٨ إلى:

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \qquad (٢\text{-}٩)$$

$$\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \qquad (٢\text{-}١٠)$$

وبحل المعادلات ٢-٩ و٢-١٠ لـ $\hat{\beta}_0$ و $\hat{\beta}_1$ نحصل على:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (٢\text{-}١١)$$

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (٢\text{-}١٢)$$

لا يتطلب تقدير المعلمات في نموذج الانحدار الخطي البسيط القائم على طريقة المربعات الصغرى أن يكون للخطأ العشوائي ε_i شكل محدد من أشكال التوزيع الاحتمالي. إذا أضفنا إلى نموذج الانحدار الخطي البسيط في المعادلة ٢-١ الافتراضي أن ε_i موزعة طبيعياً بمتوسط قيمته صفر وتباين ثابت وغير معروف قيمته σ^2، ويرمز لهذين الافتراضيين بالرمز

$N(0, \sigma^2)$، فإنه يمكن استخدام طريقة الإمكان الأكبر لتقدير المعلمات في نموذج الانحدار الخطي البسيط. الافتراض أن الأخطاء العشوائية $\varepsilon_i s$ مستقلة $N(0, \sigma^2)$ يعطي التوزيع الطبيعي لـ y_i مع:

$$E(y_i) = \beta_0 + \beta_1 x_i \quad (١٣-٢)$$

$$\operatorname{var}(y_i) = \sigma^2 \quad (١٤-٢)$$

وتكون داله الكثافة (*density function*) للتوزيع الاحتمالي الطبيعي:

$$f(y_i) = \frac{1}{\sqrt{2\,\pi\,\sigma}}\, e^{-\frac{1}{2}\left(\frac{y_i - E(y_i)}{\sigma}\right)^2} = \frac{1}{\sqrt{2\,\pi\,\sigma}}\, e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} \quad (١٥-٢)$$

نظراً لأن $y_i s$ مستقلة، فإن احتمال ملاحظة $y_n, \ldots, y_1$ هو L، والتي تمثل حاصل ضرب دوال الكثافة الفردية و$f(y_i)s$ وتمثل دالةً لكل من $\beta_0, \beta_1,$ وσ^2:

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} \frac{1}{(2\,\pi\,\sigma^2)^{1/2}}\, e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} \quad (١٦-٢)$$

إن القيم المقدّرة للمعلمات، β_0, β_1، $\hat{\sigma}^2$، والتي تُعظّم دالة الإمكان في المعادلة ٢-١٦ هي مقدّرَات الإمكان الأكبر ويمكن الحصول عليها باشتقاق دالة الإمكان بالنسبة لـ $\beta_0, \beta_1,$ وσ^2 ومساواة هذه الاشتقاقات الجزئية بالصفر. ولتسهيل الحساب، نستخدم التحويل اللوغاريتمي الطبيعي (ln) لدالة الإمكان للحصول على:

$$\frac{\partial lnL(\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}^2)}{\partial\hat{\beta}_0} = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (١٧\text{-}٢)$$

$$\frac{\partial lnL(\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}^2)}{\partial\hat{\beta}_1} = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (١٨\text{-}٢)$$

$$\frac{\partial lnL(\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}^2)}{\partial\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2}+\frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0 \quad (١٩\text{-}٢)$$

ويتم تبسيط المعادلات من ٢-١٧ إلى ٢-١٩ لتصبح:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (٢٠\text{-}٢)$$

$$\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (٢١\text{-}٢)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n} \quad (٢٢\text{-}٢)$$

المعادلتان ٢-٢٠ و٢-٢١ هما المعادلتان ٢-٩ و ٢-١٠ نفسهما. ومن ثم، فإن مقدَّرات الإمكان الأكبر لـ β_0 وβ_1 هي مقدَّرات المربعات الصغرى لـ β_0 وβ_1 نفسها المعطاة في المعادلتين ٢-١١ و٢-١٢.

وبالنسبة لنموذج الانحدار الخطي في المعادلة ٢-٢ المحتوي على متغيرات خاصية متعددة، نعرف $x_0=1$ ، ونعيد كتابة المعادلة ٢-٢ لتصبح:

$$y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \ldots + \beta_1 x_{i,p} + \varepsilon_i \qquad (٢٣-٢)$$

وبتعريف المصفوفات التالية:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

نعيد كتابة المعادلة ٢-٢٣ في شكل مصفوفة:

$$y = x\beta + \varepsilon \qquad (٢٤-٢)$$

وتكون مقدَّرَات المربعات الصغرى ومقدَّرَات الإمكان الأكبر الخاصة بالمعلمات كما يلي:

$$\hat{\beta} = (x'x)^{-1}(x'y), \qquad (٢٥-٢)$$

حيث $(x'x)^{-1}$ تمثل معكوس المصفوفة $x'x$.

الجدول (٢-١)

مجموعة بيانات الحلقات الدائرية ذات الأحمال الثقيلة مع القيمة المستهدفة المتوقعة من الانحدار الخطي

رقم الحالة **Instance**	درجة حرارة الاطلاق **Launch Temperature**	عدد الحلقات الدائرية ذات الأحمال الثقيلة **Number of O-Rings with Stress**
1	66	0
2	70	1
3	69	0
4	68	0
5	67	0
6	72	0
7	73	0
8	70	0
9	57	1
10	63	1
11	70	1
12	78	0
13	67	0
14	53	2
15	67	0
16	75	0
17	70	0
18	81	0
19	76	0
20	79	0
21	75	0
22	76	0
23	58	1

المثال (٢-١):

استخدم طريقة المربعات الصغرى لتمثيل نموذج انحدار خطي لبيانات الحلقات الدائرية في مكوك الفضاء في الجدول ١-٥، والمعطاة أيضاً في الجدول ٢-١، وقم بتحديد القيمة المستهدفة المتوقعة لكل ملحوظة باستخدام نموذج الانحدار الخطي.

تحتوي هذه البيانات على متغير خاصية واحد x يمثل درجة حرارة الإطلاق ومتغير هدف واحد y يمثل عدد الحلقات الدائرية ذات الأحمال الثقيلة. نموذج الانحدار الخطي لمجموعة البيانات هذه هو:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

يوضح الجدول ٢-٢ العملية الحسابية لتقدير $\hat{\beta}_1$ باستخدام المعادلة ٢-١١. وباستخدام المعادلة ٢-١١، نحصل على:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-65.91}{1382.82} = -0.05$$

باستخدام المعادلة ٢-١٢، نحصل على:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0.30 - (-.05)(69.57) = 3.78$$

ومن ثم، يكون نموذج الانحدار الخطي:

$$y_i = 3.78 - 0.05 x_i + \varepsilon_i$$

المعلمات في نموذج الانحدار الخطي هذا مشابهة للمعلمات $\hat{\beta}_0 = 4.301587$ و$\hat{\beta}_1 = -0.05746$ في المعادلة ١-١، والتي يتم الحصول عليها من الحزمة البرمجية إكسل لنفس مجموعة البيانات. والاختلافات الظاهرة في قيم المعلمات ناتجة عن التقريب في الحساب.

الجدول (٢-٢)

العملية الحسابية لتقدير معلمات النموذج الخطي في المثال (٢-١)

$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$y_i - \bar{y}$	$x_i - \bar{x}$	عدد الحلقات الدائرية Number of O-Rings	درجة حرارة الاطلاق Launch Temperature	رقم الحالة Instance
12.74	1.07	−0.30	−3.57	0	66	1
0.18	0.30	0.70	0.43	1	70	2
0.32	0.17	−0.30	−0.57	0	69	3
2.46	0.47	−0.30	−1.57	0	68	4
6.60	0.77	−0.30	−2.57	0	67	5
5.90	−0.73	−0.30	2.43	0	72	6
11.76	−1.03	−0.30	3.43	0	73	7
0.18	−0.13	−0.30	0.43	0	70	8
158.00	−8.80	0.70	−12.57	1	57	9
43.16	−4.60	0.70	−6.57	1	63	10
0.18	0.30	0.70	0.43	1	70	11
71.06	−2.53	−0.30	8.43	0	78	12
6.60	0.77	−0.30	−2.57	0	67	13
273.24	−28.10	1.70	−16.53	2	53	14
6.60	0.77	−0.30	−2.57	0	67	15
29.48	−1.63	−0.30	5.43	0	75	16
0.18	−0.13	−0.30	0.43	0	70	17
130.64	−3.43	−0.30	11.43	0	81	18
41.34	−1.93	−0.30	6.43	0	76	19
377.52	−5.83	−0.30	19.43	0	79	20
29.48	−1.63	−0.30	5.43	0	75	21
41.34	−1.93	−0.30	6.43	0	76	22
133.86	−8.10	0.70	−11.57	1	58	23
1382.82	−65.91			7	1600	المجموع
				$\bar{y} = 0.30$	$\bar{x} = 69.57$	المتوسط

٢-٣ نماذج الانحدار غير الخطية وتقدير المعلمة
(Nonlinear Regression Models and Parameter Estimation):

تكون نماذج الانحدار غير الخطية غير خطية في معلمات النموذج وتأخذ الشكل العام التالي:

$$y_i = f(x_i, \beta) + \varepsilon_i, \qquad (٢٦-٢)$$

حيث إن:

$$x_i = \begin{bmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,p} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

وتكون f غير خطية في β. يُعد نموذج الانحدار الأسّي التالي مثالاً على نماذج الانحدار غير الخطية:

$$y_i = \beta_0 + \beta_1 e^{\beta_2 x_i} + \varepsilon_i \qquad (٢٧-٢)$$

ويُعدُ نموذج الانحدار اللوجستي التالي مثالاً آخر على نماذج الانحدار غير الخطية:

$$y_i = \frac{\beta_0}{1 + \beta_1 e^{\beta_2 x_i}} + \varepsilon_i \qquad (٢٨-٢)$$

يتم استخدام طريقة المربعات الصغرى وطريقة الإمكان الأكبر لتقدير معلمات نموذج الانحدار غير الخطية. على عكس المعادلات ٢-٩، ٢-١٠، ٢-٢٠ و٢-٢١ لنموذج الانحدار الخطي، وبشكل عام فإن المعادلات لنموذج الانحدار غير الخطي ليس لها حلول تحليلية نظراً لأن نموذج الانحدار غير الخطي هو غير خطي في المعلمات. وتُستخدم طُرق البحث الرقمي القائمة على أسلوب البحث التكراري مثل طريقة غاوس - نيوتن (*Gauss-Newton method*) وطريقة بحث الانحدار المتدرج (*gradient decent search method*) لتحديد قيم المعلمات المقدّرة. ويمكن الحصول على شرح مفصل لطريقة غاوس-نيوتن في (*Neter et al., 1996*). وعادةً ما تُستخدم برامج حاسوبية خاصة في العديد من الحزم البرمجية الإحصائية لتقدير معلمات نموذج الانحدار غير الخطي لأنها تتطلب حسابات مكثفة لإجراء أسلوب البحث التكراري.

٢-٤ البرمجيات والتطبيقات (Software and Applications):

هناك العديد من الحزم البرمجية الإحصائية، بما في ذلك ما يلي، والتي تدعم بناء نموذج الانحدار الخطي أو غير الخطي:

- *Statistica (http://www.statsoft.com)*
- *SAS (http://www.sas.com)*
- *SPSS (http://www.ibm/com/software/analytics/spss/)*

وتُعتبر تطبيقات نماذج الانحدار الخطي وغير الخطي شائعة الاستخدام في العديد من المجالات.

التمارين (Exercises):

٢-١ بالنظر إلى مجموعة بيانات مكوك الفضاء الواردة في الجدول ٢-١، قم باستخدام المعادلة ٢-٢٥ لتقدير معلمات نموذج الانحدار الخطي التالية:

$$y_i = \beta_0 + \beta_1\sqrt{x_i} + \varepsilon_i,$$

حيث إن:

x_i هي درجة حرارة الإطلاق

y_i هي عدد الحلقات الدائرية ذات الأحمال الثقيلة

قم بحساب مجموع الأخطاء التربيعية (SSE) الناتجة عن قيم y المتوقعة من نموذج الانحدار.

٢-٢ بالنظر إلى مجموعة بيانات مكوك الفضاء الواردة في الجدول ٢-١، قم باستخدام المعادلات ٢-١١ و٢-١٢ لتقدير معلمات نموذج الانحدار الخطي التالية:

$$y_i = \beta_0 + \beta_1\sqrt{x_i} + \varepsilon_i,$$

حيث إن:

x_i هي درجة حرارة الإطلاق.

y_i هي عدد الحلقات الدائرية ذات الأحمال الثقيلة.

قم بحساب مجموع الأخطاء التربيعية (SSE) الناتجة عن قيم y المتوقعة من نموذج الانحدار.

٢-٣ قم باستخدام مجموعة البيانات الموجودة في التمرين ٢-١ لبناء نموذج الانحدار الخطي وحساب مجموع الأخطاء التربيعية (SSE) الناتجة عن قيم y المتوقعة من نموذج الانحدار.

٣- مُصنِّف بـيـيـز البسيط
Naïve Bayes Classifier

يستند مصنِّف بييز البسيط على نظرية بييز. ومن ثم، فإن هذا الفصل يستعرض أولاً نظرية بييز ثم يصف بعد ذلك مصنِّف بييز البسيط. وترد قائمة بحزم برمجية لاستكشاف البيانات التي تدعم تعلُّم مصنِّف بييز البسيط. ويتم كذلك استعراض بعض التطبيقات لمصنِّفات بييز البسيطة مع ذكر مراجعها.

٣-١ نظرية بييز (Bayes Theorem):

ليكن لدينا الحدثان A وB، يمثل تزامن أو اقتران (^) الحدثين وقوع كل من A وB في الوقت نفسه. ويتم حساب الاحتمال $P(A$^$B)$ باستخدام احتمال كل من A و B، وكل من $P(A)$ و$P(B)$، والاحتمال المشروط لـ A علماً بوقوع الحدث B،ويُكتَب $P(A|B)$، أو لـ B علماً بوقوع الحدث A، ويُكتَب $P(B|A)$:

$$P(A\text{^}B) = P(A|B)P(B) = P(B|A)P(A) \qquad (٣\text{-}١)$$

ويتم اشتقاق نظرية بييز من المعادلة ٣-١:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (٣\text{-}٢)$$

٣-٢ التصنيف القائم على نظرية بييز ومصنِّف بييز البسيط (Classification Based on the Bayes Theorem and Naïve Bayes Classifier):

بالنسبة إلى متجه البيانات x الذي يحتاج إلى تحديد فئته الهدف y، يكون التصنيف اللاحق الأكبر، (*maximum a posterior-MAP*)، y لـ x هو:

$$y_{MAP} = \arg \max_{y \in Y} P(y|x) = \arg \max_{y \in Y} \frac{p(y)P(x|y)}{P(x)} \approx \arg \max_{y \in Y} p(y)P(x|y) \quad (٣\text{-}٣)$$

حيث Y هي مجموعة كل الفئات الهدف. تُستخدم العلامة $\approx$ في المعادلة ٣-٣ لأن الاحتمال $P(x)$ هو نفسه لجميع قيم y، ومن ثم يمكن تجاهله عندما نقارن $p(y)P(x/y)/P(x)$ لجميع قيم y. $P(x)$ هو الاحتمال السابق (*prior probability*) بأننا نرصد x من دون أي معرفة عن ماهية الفئة الهدف لـ x. $P(y)$ هو الاحتمال السابق بأننا نتوقع y، مما يعكس معرفتنا المسبقة عن مجموعة البيانات لـ x وإمكانية الفئة الهدف y في مجموعة البيانات من دون الإشارة إلى أي x محددة. $P(y|x)$ هو الاحتمال اللاحق لـ y إذا علمنا أن الملحوظة المرصودة المعطاة هي x. وتقارن القيمة $\arg \max_{y \in Y} P(x|y)$ الاحتمال اللاحق لجميع الفئات الهدف بمعرفة x مسبقاً ومن ثم تختار الفئة الهدف y مع الاحتمال اللاحق الأكبر. $P(x|y)$ هو احتمال أن نرصد x إذا كانت الفئة الهدف هي y. ويكون التصنيف y الذي يعظّم $P(x|y)$ من بين جميع الفئات الهدف هو تصنيف الإمكان الأكبر (ML):

$$y_{ML} = \arg \max_{y \in Y} P(x|y) \qquad (٤\text{-}٣)$$

إذا كانت $P(y')=P(y)$ لأي $y' \neq y$ ، $Y \ni y$ ، $Y \ni y'$، فإن:

$$y_{MAP} \approx \arg \max_{y \in Y} p(y)P(x|y) \approx \arg \max_{y \in Y} P(x|y)$$

ومن ثم:

$$y_{MAP} = y_{ML}$$

ويستند مصنّف بييز البسيط على تصنيف MAP مع افتراض إضافي خاص بمتغيرات الخاصية $x=(x_1,\dots,x_p)$ أن هذه المتغيرات x_is مستقلة بعضها عن بعض. وبهذا الافتراض، يكون لدينا:

$$y_{MAP} \approx \arg \max_{y \in Y} p(y) P(x|y) = \arg \max_{y \in Y} p(y) \prod_{i=1}^{p} P(x_i|y) \quad (٣-٥)$$

ويقوم مصنِّف بييز البسيط بتقدير قيم حدود الاحتمال في المعادلة ٣-٥ على النحو التالي:

$$P(y) = \frac{n_y}{n} \quad (٣-٦)$$

$$P(x_i|y) = \frac{n_{y\&x_i}}{n_y} \quad (٣-٧)$$

حيث إن:

n	هو إجمالي عدد سجلات البيانات في مجموعة البيانات التدريبية.
n_y	هو عدد سجلات البيانات المحتوية على الفئة الهدف y.
$n_{y\&x_i}$	هو عدد سجلات البيانات بفئة الهدف y ومتغير الخاصية رقم i الذي يأخذ القيمة x_i.

المثال التالي (رقم ٣-١) يمثل تطبيقاً لمصنِّف بييز البسيط.

المثال (٣-١):

استخدم وتعرف على مصنِّف بييز البسيط لتصنيف ما إذا كان نظام تصنيع ما معطلاً باستخدام متغيرات الجودة التسعة. تعطي مجموعة البيانات التدريبية الواردة في الجدول ٣-١ جزءاً من مجموعة البيانات الواردة في الجدول ١-٤، وتتضمن تسع حالات ذات أعطال مفردة وحالة واحدة غير معطلة في نظام التصنيع. يوجد تسعة متغيرات خاصية لجودة الوحدات، (x_1, ... ,x_9)، ومتغير هدف واحد y يشير إلى عطل النظام. يوضح الجدول ٣-٢ حالات الاختبار لبعض الحالات المتعددة الأعطال.

الجدول (٣-١)

مجموعة البيانات التدريبية الخاصة بالكشف عن أعطال نظام التصنيع

متغيرات الهدف Target Variables	Attribute Variables – متغيرات الخاصية									
عطل النظام (System Fault), y	Quality of Parts – جودة وحدات المنتج									رقم الحالة Instance – (الآلة المعطلة (Faulty Machine
	x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	
1	1	0	1	0	1	0	0	0	1	1 (M1)
1	0	1	0	0	0	1	0	1	0	2(M2)
1	0	1	1	1	0	1	1	0	0	3(M3)
1	0	1	0	0	0	1	0	0	0	4(M4)
1	1	0	1	0	1	0	0	0	0	5(M5)
1	0	0	1	1	0	0	0	0	0	6(M6)
1	0	0	1	0	0	0	0	0	0	7(M7)
1	0	1	0	0	0	0	0	0	0	8(M8)
1	1	0	0	0	0	0	0	0	0	9(M9)
0	0	0	0	0	0	0	0	0	0	10(none)

باستخدام البيانات التدريبية المحددة في الجدول ٣-١، نقوم بحساب ما يلي:

$n = 10$

$n_{y=1} = 9 \quad n_{y=0} = 1$

$n_{y=1\&\,x_1=1} = 1 \quad n_{y=1\&\,x_1=0} = 8 \quad n_{y=0\,\&\,x_1=1} = 0 \quad n_{y=0\,\&\,x_1=0} = 1$

$n_{y=1\&\,x_2=1} = 1 \quad n_{y=1\&\,x_2=0} = 8 \quad n_{y=0\,\&\,x_2=1} = 0 \quad n_{y=0\,\&\,x_2=0} = 1$

$n_{y=1\&\,x_3=1} = 1 \quad n_{y=1\&\,x_3=0} = 8 \quad n_{y=0\,\&\,x_3=1} = 0 \quad n_{y=0\,\&\,x_3=0} = 1$

$n_{y=1\&\,x_4=1} = 3 \quad n_{y=1\&\,x_4=0} = 6 \quad n_{y=0\,\&\,x_4=1} = 0 \quad n_{y=0\,\&\,x_4=0} = 1$

$n_{y=1\&\,x_5=1} = 2 \quad n_{y=1\&\,x_5=0} = 7 \quad n_{y=0\,\&\,x_5=1} = 0 \quad n_{y=0\,\&\,x_5=0} = 1$

$n_{y=1\& x_6=1}=2 \quad n_{y=1\& x_6=0}=7 \quad n_{y=0\,\&\,x_6=1}=0 \quad n_{y=0\,\&\,x_6=0}=1$

$n_{y=1\& x_7=1}=5 \quad n_{y=1\& x_7=0}=4 \quad n_{y=0\,\&\,x_7=1}=0 \quad n_{y=0\,\&\,x_7=0}=1$

$n_{y=1\& x_8=1}=4 \quad n_{y=1\& x_8=0}=5 \quad n_{y=0\,\&\,x_8=1}=0 \quad n_{y=0\,\&\,x_8=0}=1$

$n_{y=1\& x_9=1}=3 \quad n_{y=1\& x_9=0}=6 \quad n_{y=0\,\&\,x_9=1}=0 \quad n_{y=0\,\&\,x_9=0}=1$

الجدول (٣-٢)

تصنيف سجلات البيانات في مجموعة البيانات التدريبية الخاصة بالكشف عن أعطال نظام التصنيع

متغير الهدف Target Variable (عطل النظام) (System Fault y		متغيرات الخاصية - Attribute Variables (جودة وحدات المنتج - Quality of Parts)									رقم الحالة Instance (الآلة المعطلة) (Faulty Machine
القيمة المصنفة (Classified Value)	القيمة الفعلية (True Value)	x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	
1	1	1	1	1	0	1	1	0	1	1	1 (M1,M2)
1	1	0	1	1	1	0	1	1	1	0	2(M2,M3)
1	1	1	0	1	1	1	0	1	0	1	3(M1,M3)
1	1	1	1	1	0	1	1	0	0	1	4(M1,M4)
1	1	1	0	1	1	1	0	0	0	1	5(M1,M6)
1	1	0	1	1	1	0	1	0	1	0	6(M2,M6)
1	1	0	1	1	0	1	1	0	1	0	7(M2,M5)
1	1	1	0	1	1	1	0	1	0	0	8(M3,M5)
1	1	0	1	1	0	0	1	0	0	0	9(M4,M7)
1	1	0	1	1	0	1	0	0	0	0	10(M5,M8)
1	1	1	1	1	1	0	1	1	0	0	11(M3,M9)
1	1	1	1	1	0	1	0	0	0	1	12(M1,M8)
1	1	1	1	1	1	1	1	1	1	1	13(M1,M2,M3)
1	1	1	1	1	1	1	1	1	1	0	14(M2,M3,M5)
1	1	1	1	1	1	0	1	1	1	0	15(M2,M3,M9)
1	1	1	1	1	1	1	0	0	0	1	16(M1,M6,M8)

ويتم تصنيف الحالة أو السجل رقم ١ في الجدول ٣-١ بقيم لـ x=(1, 0, 0, 0, 1, 0, 1,0,1) على النحو التالي:

$$p(y=1)\prod_{i=1}^{9}P(x_i|y=1)=\frac{n_{y=1}}{n}\prod_{i=1}^{9}\frac{n_{y=1\,\&\,x_i}}{n_{y=1}}$$

$$=\frac{n_{y=1}}{n}\Big(\frac{n_{y=1\,\&\,x_1=1}}{n_{y=1}}\times\frac{n_{y=1\,\&\,x_2=0}}{n_{y=1}}\times\frac{n_{y=1\,\&\,x_3=0}}{n_{y=1}}\times\frac{n_{y=1\,\&\,x_4=0}}{n_{y=1}}$$

$$\times\frac{n_{y=1\,\&\,x_5=1}}{n_{y=1}}\times\frac{n_{y=1\,\&\,x_6=0}}{n_{y=1}}\times\frac{n_{y=1\,\&\,x_7=1}}{n_{y=1}}$$

$$\times\frac{n_{y=1\,\&\,x_8=0}}{n_{y=1}}\times\frac{n_{y=1\,\&\,x_9=1}}{n_{y=1}}\Big)$$

$$=\frac{9}{10}\left(\frac{1}{9}\times\frac{8}{9}\times\frac{8}{9}\times\frac{6}{9}\times\frac{2}{9}\times\frac{7}{9}\times\frac{5}{9}\times\frac{5}{9}\times\frac{3}{9}\right)>\ 0$$

$$p(y=0)\prod_{i=1}^{9}P(x_i|y=0)=\frac{n_{y=0}}{n}\prod_{i=1}^{9}\frac{n_{y=0\,\&\,x_i}}{n_{y=0}}$$

$$=\frac{n_{y=0}}{n}\Big(\frac{n_{y=0\,\&\,x_1=1}}{n_{y=0}}\times\frac{n_{y=0\,\&\,x_2=0}}{n_{y=0}}\times\frac{n_{y=0\,\&\,x_3=0}}{n_{y=0}}\times\frac{n_{y=0\,\&\,x_4=0}}{n_{y=0}}$$

$$\times\frac{n_{y=0\,\&\,x_5=1}}{n_{y=0}}\times\frac{n_{y=0\,\&\,x_6=0}}{n_{y=0}}\times\frac{n_{y=0\,\&\,x_7=1}}{n_{y=0}}$$

$$\times\frac{n_{y=0\,\&\,x_8=0}}{n_{y=0}}\times\frac{n_{y=0\,\&\,x_9=1}}{n_{y=0}}\Big)$$

$$=\frac{1}{10}\left(\frac{0}{1}\times\frac{1}{1}\times\frac{1}{1}\times\frac{1}{1}\times\frac{0}{1}\times\frac{1}{1}\times\frac{0}{1}\times\frac{1}{1}\times\frac{0}{1}\right)=\ 0$$

$$y_{MAP} \approx \arg \max_{y \in Y} p(y) \prod_{i=1}^{p} P(x_i|y) = 1$$

(هذه النتيجة تعني أن النظام به أعطال)

يمكن تصنيف الحالات من رقم ٢ إلى ٩ في الجدول ٣-١ وجميع الحالات في الجدول ٣-٢ على نحو مماثل للحصول على $y_{MAP} = 1$ لأنه يوجد $x_i = 1$ و $n_{y=0\&xi=1} / n_{y=0} = 0/1$ ، مما يجعل $p(y = 0)P(x|y = 0) = 0$. يتم تصنيف الحالة رقم ١٠ في الجدول ٣-١ بالقيم $x = (0, 0, 0, 0, 0, 0, 0, 0, 0)$ على النحو التالي:

$$y_{MAP} \approx \arg \max_{y \in Y} p(y) \prod_{i=1}^{p} P(x_i|y) = 0$$

(هذه النتيجة تعني أن النظام ليس به أعطال)

ومن ثم، يتم تصنيف جميع الحالات في الجدولين ٣-١ و٣-٢ بشكل صحيح بواسطة مصنّف بييز البسيط.

٣-٣ البرمجيات والتطبيقات (Software and Applications):

تدعم حزم البرمجيات التالية تعلم مصنّف بييز البسيط:

- *Weka (http://www.cs.waikato.ac.nz/ml/weka/)*
- *MATLAB ® (http://www.mathworks.com)*

ولقد تمّ تطبيق مصنّف بييز البسيط بنجاح في العديد من المجالات، بما في ذلك تصنيف النصوص والوثائق، والموجود على الرابط:

(http://www.cs.waikato.ac.nz/~eibe/pubs/FrankAndBouckaertPKDD 06new.pdf)

التمارين (Exercises):

٣-١ قم ببناء مصنّف بييز البسيط لتصنيف المتغير الهدف من متغير الخاصية في بيانات البالون (*Balloon data set*) الواردة في الجدول ١-١، ومن ثُمّ تقييم أداء التصنيف لمصنّف بييز البسيط من خلال حساب ما هي النسبة المئوية لسجلات البيانات في مجموعة البيانات التي يتم تصنيفها بشكل صحيح بواسطة مصنّف بييز البسيط.

٣-٢ في بيانات الحلقات الدائرية في مكوك الفضاء (*Space shuttle O-rings data set*) الواردة في الجدول ١-٢، افترض أن متغير الخاصية ضغط التحقق من التسرب (*leak-check pressure*) كخاصية نوعية ذات ثلاث قيم نوعية، وأن عدد الحلقات الدائرية ذات الأحمال الثقيلة (*number of O-rings with stress*) كمتغير هدف نوعي ذي ثلاث قيم نوعية. قم ببناء مصنّف بييز البسيط لتصنيف متغير الهدف: الحلقات الدائرية ذات الأحمال الثقيلة، من متغير الخاصية: ضغط التحقق من التسرب ومن ثم قم بتقييم أداء تصنيف مصنّف بييز البسيط من خلال حساب النسبة المئوية لسجلات البيانات في مجموعة البيانات التي يتم تصنيفها بشكل صحيح بواسطة مصنّف بييز البسيط.

٣-٣ قم ببناء مصنّف بييز البسيط لتصنيف المتغير الهدف من متغيرات الخاصية في مجموعة بيانات العدسات (*lenses data set*) المحددة في الجدول ١-٣، ومن ثم قم بتقييم أداء تصنيف مصنّف بييز البسيط من خلال حساب النسبة المئوية لسجلات البيانات في مجموعة البيانات التي يتم تصنيفها بشكل صحيح بواسطة مصنّف بييز البسيط.

٣-٤ قم ببناء مصنّف بييز البسيط لتصنيف المتغير الهدف من متغيرات الخاصية في مجموعة البيانات الموجودة في التمرين ١-١، ومن ثمّ قم بتقييم أداء تصنيف مصنّف بييز البسيط من خلال حساب النسبة المئوية لسجلات البيانات في مجموعة البيانات التي يتم تصنيفها بشكل صحيح بواسطة مصنّف بييز البسيط.

٤- أشجار القرار والانحدار
Decision and Regression Trees

تُستخدَم أشجار القرار والانحدار للتعرف على أنماط التصنيف والتنبؤ من البيانات، والتعبير عن العلاقة بين متغيرات الخاصية x مع المتغير الهدف، y، $y = F(x)$، على شكل شجرة. تقوم شجرة القرار بتصنيف قيمة الهدف النوعي لسجل بيانات باستخدام قيم الخاصية الخاصة بها. بينما تتنبأ شجرة الانحدار بقيمة الهدف الرقمية لسجل بيانات باستخدام قيم الخاصية الخاصة بها.

في هذا الفصل، سنقوم أولاً بتعريف شجرة القرار الثنائية، وسنتناول أيضاً الخوارزمية التي تقوم بمعرفة وتعلم شجرة قرار ثنائية من مجموعة بيانات ذات متغيرات خاصية نوعية عديدة ومتغير هدف نوعي واحد. ثم يتم وصف طريقة التعرف على وتعلم شجرة القرار غير الثنائية. وسيتم التطرق إلى مفاهيم إضافية للتعامل مع متغيرات الخاصية الرقمية، والقيم المفقودة لمتغيرات الخاصية، والتعامل مع متغير الهدف الرقمي لبناء شجرة الانحدار. وسيتم استعراض قائمة بحزم برمجية لاستكشاف البيانات التي تدعم تعلُّم أشجار القرار والانحدار. سيتم أيضاً استعراض بعض التطبيقات الخاصة بأشجار القرار والانحدار مع ذكر مراجعها.

٤-١ تعلُّم شجرة القرار الثنائية وتصنيف البيانات باستخدام شجرة القرار (Learning a Binary Decision Tree and Classifying Data Using a Decision Tree):

في هذا الجزء، يتم استعراض عناصر شجرة القرار، وتقوم دوال انتقاء الانفصال (*split selection methods*) بتقديم الأساس المنطقي لبناء شجرة قرار ذات وصف طوله يكون بالحد الأدنى. أخيراً، سيتم توضيح كيفية بناء شجرة قرار من الأعلى إلى الأسفل.

٤-١-١ عناصر شجرة القرار (Elements of a Decision Tree):

يبين الجدول ٤-١ جزءاً من مجموعة البيانات لنظام تصنيع ما والموضحة بشكل كامل في الجدول ١-٤. حيث تتضمن مجموعة البيانات في الجدول ٤-١ تسعة من متغيرات الخاصية لجودة وحدات المنتج، ومتغير هدف واحد يوضح ما إذا كان النظام معطلاً أم لا. يتم استخدام مجموعة البيانات هذه كمجموعة بيانات تدريبية لاستخلاص شجرة قرار ثنائية لتصنيف ما إذا كان النظام معطلاً أم لا باستخدام قيم متغيرات الجودة التسعة. ويبين الشكل ٤-١ شجرة

القرار الثنائية الناتجة لتوضيح عناصر شجرة القرار. وسوف يتم توضيح الكيفية التي تمّ بها استخلاص شجرة القرار هذه في مكان آخر. وكما هو مبين في الشكل ٤-١، فإن شجرة القرار الثنائية عبارة عن رسم بياني ذي عدة عُقَد (*nodes*). حيث تقع عقدة الجذر (*root node*) في أعلى الشجرة وتتكون هذه العقدة من جميع سجلات البيانات في مجموعة البيانات التدريبية.

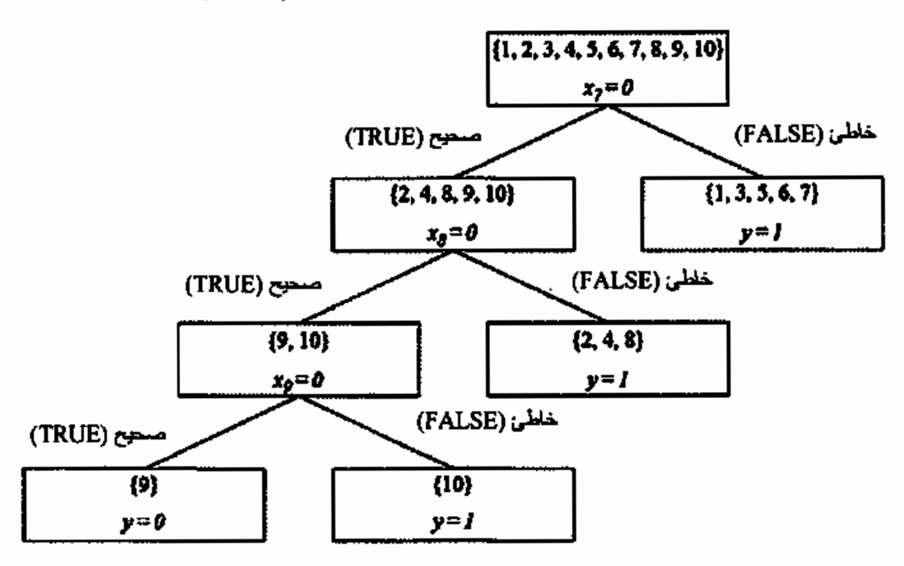
بالنسبة لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع، تحتوي عقدة الجذر على مجموعة مكونة من كل سجلات البيانات العشرة في مجموعة البيانات التدريبية، {1، 2،...، 10}. لاحظ أن الأرقام في مجموعة البيانات هي أرقام لكل حالة على حدة. يتم فصل السجلات الموجودة في عقد الجزر إلى مجموعتين فرعيتين، {2، 4، 8، 9، 10 } و{1، 3، 5، 6، 7}، وذلك باستخدام متغير الخاصية، x_7 واثنين من القيم النوعية لهذا المتغير، $x_7 = 0$ و $x_7 = 1$ جميع الحالات في المجموعة الفرعية، {2، 4، 8، 9، 10}، تكون بها قيمة $x_7 = 0$ وجميع الحالات في المجموعة الفرعية، {1، 3، 5، 6، 7}، تكون بها قيمة $x_7 = 1$ يتم تمثيل كل مجموعة فرعية كعقدة في شجرة القرار.

الجدول (٤-١)

مجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع

متغيرات الهدف Target Variables	Attribute Variables – متغيرات الخاصية									
عطل النظام (System Fault), y	Quality of Parts – جودة وحدات المنتج									رقم الحالة Instance (الآلة المعطلة – (Faulty Machine
	x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	
1	1	0	1	0	1	0	0	0	1	1 (M1)
1	0	1	0	0	0	1	0	1	0	2(M2)
1	0	1	1	1	0	1	1	0	0	3(M3)
1	0	1	0	0	0	1	0	0	0	4(M4)
1	1	0	1	0	1	0	0	0	0	5(M5)
1	0	0	1	1	0	0	0	0	0	6(M6)
1	0	0	1	0	0	0	0	0	0	7(M7)
1	0	1	0	0	0	0	0	0	0	8(M8)
1	1	0	0	0	0	0	0	0	0	9(M9)
0	0	0	0	0	0	0	0	0	0	10(none)

الشكل (٤-١)
شجرة القرار الخاصة بالكشف عن أعطال نظام التصنيع



ويُستخدم التعبير المنطقي في شجرة القرار للتعبير عن $x_7 = 0$ باستخدام $x_7= 0$ كتعبير منطقيٍ صحيح (*TRUE*)، و$x_7 = 1$ باستخدام $x_7= 0$ كتعبير منطقي خاطئ (*FALSE*). ويُسمى $x_7 = 0$ بشرط الانقسام أو الانفصال (معيار الانقسام أو الفصل)، وقيمها الصحيحة (*TRUE*) والخاطئة (*FALSE*) تسمح بانقسام ثنائي لمجموعة السجلات في عقدة الجذر إلى فرعين بوجود عقدة في نهاية كل فرع. كل من العقدتين الجديدتين يمكن أن تنقسم إلى مزيد من العقد باستخدام أحد متغيرات الخاصية المتبقية في معيار الانقسام، أو الفصل. ولا يمكن تقسيم عقدة ما مرةً أخرى إذا كانت سجلات البيانات في مجموعة البيانات في هذه العقدة لها قيمة المتغير الهدف نفسه. وتصبح هذه العقدة عندئذ عقدة ورقة (*leaf node*) في شجرة القرار. وباستثناء عقدة الجذر وعقدة الورقة، فإن العقد الأخرى في شجرة القرار تسمى العقد الداخلية (*internal nodes*).

يمكن لشجرة القرار أن تُصنِّف سجل بيانات معيّناً عن طريق تمرير سجل البيانات من خلال شجرة القرار باستخدام قيم متغيرات الخاصية في سجل البيانات. على سبيل المثال، يتم فحص سجل البيانات للحالة رقم ١٠ أولاً مع شرط الانفصال الأول في عقدة الجذر. وحيث

إن $x_7 = 0$ يتم تمرير سجل البيانات إلى الفرع الأيسر من الشجرة. وحيث إن $x_8 = 0$ ومن ثمّ $x_9 = 0$ يتم تمرير سجل البيانات وصولاً إلى عقدة الورقة أقصى اليسار. ويأخذ سجل البيانات القيمة الهدف لعقدة الورقة تلك، $y = 0$، والذي يصنف سجل البيانات على أنه نظام غير معطل.

٤-١-٢ شجرة القرار ذات طول الوصف الأصغر (Decision Tree with the Minimum Description Length):

ابتداء من عقدة الجذر المحتوية على جميع سجلات البيانات في مجموعة البيانات التدريبية، هناك تسع طرق ممكنة لتقسيم عقدة الجذر باستخدام متغيرات الخاصية التسعة بشكل فردي في شرط الانفصال. ولكل عقدة في نهاية فرع الشجرة بعد انقسام عقدة الجذر، يوجد ثماني طرق ممكنة لتقسيم العقدة باستخدام كل من متغيرات الخاصية الثمانية المتبقية بشكل فردي.

وتستمر هذه العملية، ويمكن أن ينتج عنها العديد من أشجار القرار الممكنة. كل أشجار القرار الممكنة تختلف في حجمها وتعقيدها. يمكن لشجرة القرار أن تكون كبيرة بحيث يكون لديها عدد من عقد الأوراق مساوياً لسجلات البيانات في مجموعة البيانات التدريبية بحيث تكون كل عقدة ورقة محتوية على سجل بيانات واحد ويمكن أن نتساءل. أي أشجار القرار الممكنة ينبغي أن يُستخدَم لتمثيل F وهي العلاقة بين متغيرات الخاصية مع متغير الهدف؟ تهدف خوارزمية شجرة القرار إلى الحصول على أصغر شجرة القرار التي يمكنها تمثيل F، وهو ما يعني، شجرة القرار التي تتطلب الحد الأدنى من طول الوصف (وتُسمى شجرة القرار ذات طول الوصف الأصغر). بافتراض أن لدينا كلاً من شجرة القرار الصغرى وشجرة القرار الكبرى التي تصنف جميع سجلات البيانات في مجموعة البيانات التدريبية بشكل صحيح، فمن المتوقع أن شجرة القرار الصغرى تُعمّم أنماط التصنيف بشكل أفضل من شجرة القرار الكبرى، وأن أنماط التصنيف الأفضل والمُعمّمة تسمح بتصنيف أفضل لمزيد من نقاط البيانات بما في ذلك نقاط البيانات غير الموجودة في مجموعة البيانات التدريبية. لنفترض أن لدينا شجرة قرار كبيرة بها عدد من عقد الأوراق مساوٍ لسجلات البيانات في مجموعة البيانات التدريبية بحيث تكون كل عقدة ورقة محتوية على سجل بيانات واحد. على الرغم من أن شجرة القرار الكبيرة هذه تقوم بتصنيف كافة سجلات البيانات التدريبية بشكل صحيح، إلا

أن أداءها قد يكون ضعيفاً عند تصنيف سجلات بيانات جديدة غير موجودة في مجموعة البيانات التدريبية.

ويعود ذلك إلى أن سجلات البيانات الجديدة هذه تحتوي على مجموعة قيم جديدة ومختلفة عن تلك الموجودة في سجلات البيانات في مجموعة البيانات التدريبية، ومن ثم لا تسلك نفس مسارات سجلات البيانات وصولاً إلى عقد الورقة في شجرة القرار. نحن بحاجة إلى شجرة قرار تقوم بتمثيل أنماط تصنيف معممة للعلاقة F. كلما زاد مستوى التعميم للعلاقة F، قصر طول الوصف الخاص بها، لأنها تخفي الاختلافات البسيطة بين سجلات البيانات الفردية. ومن ثم، كلما صغرت شجرة القرار، كبرت قدرة التعميم لشجرة القرار كما هو متوقع لها أن تكون.

٤-١-٣ طرق انتقاء الانفصال (Split Selection Methods):

سعياً إلى شجرة قرار ذات حد أدنى لطول الوصف، نحتاج إلى معرفة كيفية انقسام أو فصل عقدة ما حتى نتمكن من تحقيق الهدف المتمثل في الحصول على شجرة القرار ذات حد أدنى لطول الوصف. لنأخذ مثالاً يوضح كيفية بناء شجرة قرار من مجموعة البيانات في الجدول ٤-١. هناك تسع من الطرق الممكنة لفصل عقدة الجذر باستخدام متغيرات الخاصية التسعة بشكل فردي، كما هو مبين في الجدول ٤-٢.

أيٍ معايير الانقسام أو الانفصال التسعة يتوجب استخدامه لكي نحصل على أصغر شجرة قرار؟ النهج المتعارف عليه لانتقاء طريقة الانفصال هو اختيار الانفصال الذي ينتج عنه مجموعات بيانات فرعية أكثر تجانساً. مجموعة البيانات المتجانسة هي مجموعة البيانات التي يكون لسجلاتها قيمة متغير الهدف نفسه. يوجد مقاييس متنوعة يتم استخدامها لقياس تجانس البيانات مثل: مقياس عشوائية المعلومات *(Information entropy)*، ومؤشر جيني *(gini – index)*، إلخ *(Breiman et al., 1984;Quinlan, 1986; Ye, 2003)*.

يتم استخدام مقياس عشوائية المعلومات بشكل أساسي لقياس عدد بتات *(bits)*، أو خوينات، المعلومات اللازمة لتشفير البيانات. يتم تعريف عشوائية المعلومات كما يلي:

$$\text{entropy}(D) = \sum_{\mathrm{i}=1}^{c} -P_{\mathrm{i}} \log_2 P_{\mathrm{i}} \qquad (١-٤)$$

$$-0 \log_2 0 = 0 \qquad (٢-٤)$$

$$\sum_{\mathrm{i}=1}^{c} P_{\mathrm{i}} = 1 \qquad (٣-٤)$$

حيث إن:

D تشير إلى مجموعة البيانات المُعطاة.

c تشير إلى عدد قيم الهدف المختلفة.

P_i تشير إلى احتمال أن سجل بيانات معين في مجموعة البيانات يأخذ قيمة الهدف i .

الجدول (٤-٢) الانفصال الثنائي لعقدة الجذر والعملية الحسابية لقيمة مقياس عشوائية المعلومات لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع

المجموعات الفرعية الناتجة ومتوسط مقياس عشوائية المعلومات للانفصال Resulting Subsets and Average Information Entropy of Split	شرط الانفصال أو الانقسام Split Criterion
{2, 3, 4, 5, 6, 7, 8, 9, 10}, {1} $entropy(S)=\frac{9}{10}entropy(D_{true})+\frac{1}{10}entropy(D_{false})$ $=\frac{9}{10}\times\left(-\frac{8}{9}\log_2\frac{8}{9}-\frac{1}{9}\log_2\frac{1}{9}\right)+\frac{1}{10}\times 0=0.45$	$x_1=0$: TRUE or FALSE
{1, 3, 4, 5, 6, 7, 8, 9, 10}, {2} $entropy(S)=\frac{9}{10}entropy(D_{true})+\frac{1}{10}entropy(D_{false})$ $=\frac{9}{10}\times\left(-\frac{8}{9}\log_2\frac{8}{9}-\frac{1}{9}\log_2\frac{1}{9}\right)+\frac{1}{10}\times 0=0.45$	$x_2=0$: TRUE or FALSE
{1, 2, 4, 5, 6, 7, 8, 9, 10}, {3} $entropy(S)=\frac{9}{10}entropy(D_{true})+\frac{1}{10}entropy(D_{false})$ $=\frac{9}{10}\times\left(-\frac{8}{9}\log_2\frac{8}{9}-\frac{1}{9}\log_2\frac{1}{9}\right)+\frac{1}{10}\times 0=0.45$	$x_3=0$: TRUE or FALSE
{1, 5, 6, 7, 8, 9, 10}, {2, 3, 4} $entropy(S)=\frac{7}{10}entropy(D_{true})+\frac{3}{10}entropy(D_{false})$ $=\frac{7}{10}\times\left(-\frac{6}{7}\log_2\frac{6}{7}-\frac{1}{7}\log_2\frac{1}{7}\right)+\frac{3}{10}\times 0=0.41$	$x_4=0$: TRUE or FALSE
{2, 3, 4, 6, 7, 8, 9, 10}, {1, 5} $entropy(S)=\frac{8}{10}entropy(D_{true})+\frac{2}{10}entropy(D_{false})$ $=\frac{8}{10}\times\left(-\frac{7}{8}\log_2\frac{7}{8}-\frac{1}{8}\log_2\frac{1}{8}\right)+\frac{2}{10}\times 0=0.43$	$x_5=0$: TRUE or FALSE
{1, 2, 4, 5, 7, 8, 9, 10}, {3, 6} $entropy(S)=\frac{8}{10}entropy(D_{true})+\frac{2}{10}entropy(D_{false})$ $=\frac{8}{10}\times\left(-\frac{7}{8}\log_2\frac{7}{8}-\frac{1}{8}\log_2\frac{1}{8}\right)+\frac{2}{10}\times 0=0.43$	$x_6=0$: TRUE or FALSE
{2, 4, 8, 9, 10}, {1, 3, 5, 6, 7} $entropy(S)=\frac{5}{10}entropy(D_{true})+\frac{5}{10}entropy(D_{false})$ $=\frac{5}{10}\times\left(-\frac{4}{5}\log_2\frac{4}{5}-\frac{1}{5}\log_2\frac{1}{5}\right)+\frac{5}{10}\times 0=0.36$	$x_7=0$: TRUE or FALSE

يتبع

تابع الجدول (٤-٢) الانفصال الثاني لعقدة الجذر والعملية الحسابية لقيمة مقياس عشوائية المعلومات لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع

شرط الانفصال أو الانقسام Split Criterion	المجموعات الفرعية الناتجة ومتوسط مقياس عشوائية المعلومات للانفصال Resulting Subsets and Average Information Entropy of Split
$x_8 = 0$: TRUE or FALSE	{1, 5, 6, 7, 9, 10}, {2, 3, 4, 8} $\text{entropy}(S)=\frac{6}{10}\text{entropy}(D_{true})+\frac{4}{10}\text{entropy}(D_{false})$ $=\frac{6}{10}\times\left(-\frac{5}{6}\log_2\frac{5}{6}-\frac{1}{6}\log_2\frac{1}{6}\right)+\frac{4}{10}\times 0=0.39$
$x_9 = 0$: TRUE or FALSE	{2, 3, 4, 6, 7, 8, 10}, {1, 5, 9} $\text{entropy}(S)=\frac{7}{10}\text{entropy}(D_{true})+\frac{3}{10}\text{entropy}(D_{false})$ $=\frac{7}{10}\times\left(-\frac{6}{7}\log_2\frac{6}{7}-\frac{1}{7}\log_2\frac{1}{7}\right)+\frac{3}{10}\times 0=0.41$

تقع قيمة العشوائية (*entropy value*) في النطاق $[0,\log_2 c]$. على سبيل المثال، في مجموعة البيانات في الجدول ٤-١، لدينا $c = 2$ (لقيمتي الهدف، $y = 0$، و$y = 1$)، $P_1 = 9/10=0.9$ (٩ من ١٠ سجلات بها قيمة الهدف $y=0$)، $P_2 = 1/10=0.1$ (١ من ١٠ سجلات بها قيمة الهدف $y=1$)، و

$$\text{entropy}(D) = \sum_{i=1}^{2} -P_i \log_2 P_i = -0.9 \log_2 0.9 - 0.1 \log_2 0.1 = 0.47.$$

يوضح الشكل ٤-٢ كيف أن قيمة عشوائية المعلومات تتغير مع P_1 ($P_2 = 1 - P_1$) عندما تكون $c=2$. وبصورة خاصة، يكون لدينا:

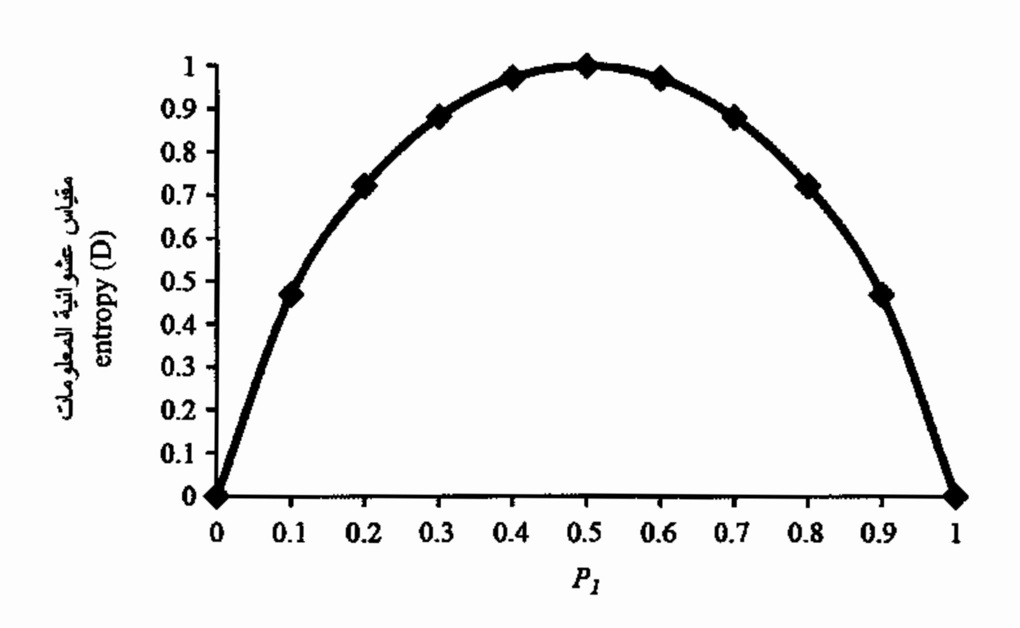
- $P_1= 0.5, P_2=0.5$, entropy $(D) = 1$
- $P_1= 0, P_2=1$, entropy $(D) = 0$
- $P_1= 1, P_2=0$, entropy $(D) = 0$

إذا كانت كل سجلات البيانات في مجموعة البيانات تأخذ قيمة الهدف نفسها، يكون لدينا $P_1=0$ ، $P_2=1$ أو $P_1=1$، $P_2=0$، وتكون قيمة عشوائية المعلومات هي صفر، وهو ما يعني، أننا بحاجة إلى عدد صفر من بتات ($bits$)، أو خوينات، المعلومات لأننا نعرف مسبقاً قيم الهدف الذي اتخذته جميع سجلات البيانات. ومن ثم، فإن قيمة عشوائية المعلومات المساوية للصفر تشير إلى أن مجموعة البيانات متجانسة فيما يخص قيمة متغير الهدف. إذا كان لنصف مجموعة واحدة من سجلات البيانات نفس قيمة متغير الهدف، وللنصف الآخر من المجموعة قيمة هدف أخرى، يكون لدينا $P_1=0.5$، $P_2=0.5$، وتكون قيمة عشوائية المعلومات هي 1، وهذا يعني أننا نحتاج إلى عدد بِت واحد (أو خوينة واحدة) من المعلومات لإيجاد قيمة الهدف. ومن ثم، فإن قيمة عشوائية المعلومات تشير إلى أن مجموعة البيانات غير متجانسة. عندما نستخدم مقياس عشوائية المعلومات لقياس تجانس البيانات، فإنه كلما انخفضت قيمة عشوائية المعلومات، تجانست مجموعة البيانات بالنسبة لقيمة متغير الهدف.

بعد انفصال مجموعة البيانات إلى عدة مجموعات فرعية، يتم استخدام المعادلة التالية لحساب قيمة متوسط عشوائية المعلومات للمجموعات الفرعية:

$$\text{entropy}(S) = \sum_{v \in Values(S)} \frac{|D_v|}{|D|} \text{entropy}(D_v) \qquad (٤\text{-}٤)$$

الشكل (٤-٢)
عشوائية المعلومات



حيث:

S	تشير إلى الانفصال.
$Values\ (S)$	تشير إلى مجموعة القيم التي يتم استخدامها في الانفصال
v	تشير إلى قيمة موجودة في $Values\ (S)$.
D	تشير إلى مجموعة البيانات التي يتم فصلها.
$\|D\|$	تشير إلى عدد سجلات البيانات في مجموعة البيانات D.
D_v	تشير إلى المجموعة الفرعية الناتجة عن الانفصال باستخدام قيمة الانفصال v.
$\|D_v\|$	تشير إلى عدد سجلات البيانات في مجموعة البيانات D_v.

على سبيل المثال، عقدة الجذر لشجرة قرار مجموعة البيانات في الجدول ٤-١ لها مجموعة البيانات $D = \{1, 2, ..., 10\}$ حيث قيمة عشوائية المعلومات تساوي ٠,٤٧، كما هو

موضح سابقاً. باستخدام معيار الانفصال، $x_1 = 0$: *(TRUE)* أو *(FALSE)*، يتم تقسيم عقدة الجذر إلى قسمين فرعيين: القسم الأول $D_{false} = \{1\}$ وهو متجانس، والقسم الثاني $D_{true} = \{2,3,4,5,6,7,8,9,10\}$ وهو غير متجانس بوجود ثمانية سجلات قيمة الهدف لها واحد، وسجل بيانات واحد يأخذ قيمة الهدف صفر. متوسط عشوائية المعلومات للمجموعات الفرعية الاثنتين بعد الانفصال هو:

$$\text{entropy}(S) = \frac{9}{10}\text{entropy}(D_{true}) + \frac{1}{10}\text{entropy}(D_{false})$$
$$= \frac{9}{10} \times \left(-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9}\right) + \frac{1}{10} \times 0 = 0.45.$$

حيث إن قيمة متوسط عشوائية المعلومات للمجموعات الفرعية بعد الانفصال أفضل من قيمة عشوائية المعلومات لـ $(D)=0.47$، فإن هذا الانفصال يُحسّن من تجانس البيانات. يوضح الجدول ٤-٢ متوسط عشوائية المعلومات للمجموعات الفرعية بعد إجراء كل من الانفصاليات الثمانية الأخرى لعقدة الجذر. من بين الانفصالات التسعة الممكنة، فإن الانفصال الذي يستخدم المعيار $x_7 = 0$ *(TRUE)* أو *(FALSE)* ينتج عنه المتوسط الأقل لعشوائية المعلومات، مما يدل على مجموعات فرعية أكثر تجانساً. ومن ثم، فإن معيار الانفصال $x_7 = 0$ *(TRUE)* أو*(FALSE)* يتم اختياره لفصل عقدة الجذر، مما ينتج عنه عقدتان داخليتان كما هو مبين في الشكل ٤-١. العقدة الداخلية مع المجموعة الفرعية، {2، 4، 8، 9، 10}، ليست متجانسة. ومن ثم، تتفرع شجرة القرار هذه إلى المزيد من الانفصالات حتى تصبح جميع عقد الأوراق متجانسة.

يتم تعريف مؤشر جيني *(gini – index)*، مقياس آخر لتجانس البيانات، على النحو التالي:

$$\text{gini}(D) = 1 - \sum_{i=1}^{c} P_i^2 \qquad (٤\text{-}٥)$$

على سبيل المثال، وباستخدام مجموعة البيانات المعطاة في الجدول ٤-١، يكون لدينا $C=2$، $P_1 = 0.9$، $P_2=0.1$ و

$$\text{gini}(D) = 1 - \sum_{i=1}^{c} P_i^2 = 1 - 0.9^2 - 0.1^2 = 0.18$$

يتم احتساب قيم مؤشر جيني لـ $c=2$ والقيم التالية لـ P_i:

- $P_1= 0.5, P_2= 0.5, gini\ (D)=1-0.5^2-0.5^2=0.5$
- $P_1= 0, P_2= 1, gini\ (D)=1- 0^2- 1^2= 0$
- $P_1= 1, P_2= 0, gini\ (D)=1- 1^2- 0^2= 0$

ومن ثم، كلما صغُرت قيمة مؤشر جيني، كانت مجموعة البيانات أكثر تجانساً. يتم حساب متوسط قيمة مؤشر جيني للمجموعات الفرعية للبيانات بعد الانفصال، على النحو التالي:

$$\text{gini}(S) = \sum_{v \in Values(S)} \frac{|D_v|}{|D|} \text{gini}(D_v) \qquad (٤-٦)$$

يوضح الجدول ٤-٣ متوسط قيمة مؤشر جيني للمجموعات الفرعية بعد إجراء كل من الانفصالات التسعة لعقدة الجذر لمجموعة البيانات التدريبية الخاصة بالكشف عن الأعطال بنظام التصنيع. من بين التسعة انفصالات المحتملة، فإن معيار الانفصال لـ $x_7 = 0$: *(TRUE)* أو *(FALSE)* ينتج عنه أصغر قيمة لمتوسط مؤشر جيني، والذي يشير إلى المجموعات الفرعية الأكثر تجانساً. يتم اختيار معيار الانفصال $x_7 = 0$: *(TRUE)* أو *(FALSE)* لفصل عقدة الجذر. ومن ثم، فإن استخدام مؤشر جيني قد نتج عنه الانفصال نفسه المستخدم مع مقياس عشوائية المعلومات.

٤-١-٤ خوارزمية بناء شجرة القرار من أعلى إلى أسفل
(Algorithm for the Top-Down Construction of a Decision Tree):

يصف هذا الجزء ويوضح خوارزمية بناء شجرة قرار كاملة. تكون خطوات خوارزمية بناء شجرة القرار الثنائية (البناء من أعلى إلى أسفل) كالتالي:

١- ابدأ من عقدة الجذر التي تشتمل على جميع سجلات البيانات في مجموعة البيانات التدريبية واختر هذه العقدة لإجراء الانفصال.

٢- قم بتطبيق دالة انتقاء الانفصال للعقدة المختارة لتحديد أفضل انفصال والذي يتماشى مع معيار الانفصال، ثم قم بتقسيم مجموعة سجلات البيانات التدريبية الموجودة في العقدة المختارة إلى عقدتين مع مجموعتين فرعيتين لسجلات البيانات، على التوالي.

٣- افحص ما إذا كان معيار التوقف عن التكرار قد تحقق. إذا كان الأمر كذلك، يكون قد اكتمل بناء الشجرة؛ خلاف ذلك، يتم العودة إلى الخطوة ٢ للاستمرار في اختيار عقدة أخرى يتم فصلها.

يقوم معيار التوقف عن التكرار والمبني على أساس تجانس البيانات بإيقاف التكرار في الخوارزمية عندما يكون لدى كل عقدة من عقد الورقة بيانات متجانسة، وهو ما يعني، مجموعة سجلات البيانات ذات نفس القيمة الهدف. إنّ العديد من مجموعات البيانات الكبيرة والحقيقية عادةً ما تكون مشوشة وغير نقية، مما يجعل الأمر صعباً للحصول على مجموعة بيانات متجانسة في عقد الورقة. ومن ثم، غالباً ما يتم ربط معيار التوقف عن التكرار في الخوارزمية بمقياس لتجانس البيانات ليكون المعيار أصغر من قيمة محددة، على سبيل المثال، يتم التوقف عن التكرار عندما يكون مقياس عشوائية المعلومات أقل من $(entropy(D) < 0.1)$. فيما يلي سيتم توضيح كيفية بناء شجرة قرار ثنائية كاملة لمجموعة بيانات الكشف عن أعطال نظام التصنيع.

الجدول (٤-٣) الانفصال الثنائي لعقدة الجذر والعملية الحسابية لقيمة مؤشر جيني لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع

شرط الانفصال أو الانقسام **Split Criterion**	المجموعات الفرعية الناتجة ومتوسط قيمة مؤشر جيني للانفصال **Resulting Subsets and Average Gini-Index Value of Split**
$x_1 = 0$: TRUE or FALSE	{2, 3, 4, 5, 6, 7, 8, 9, 10}, {1} $\text{gini}(S) = \frac{9}{10}\text{gini}(D_{true}) + \frac{1}{10}\text{gini}(D_{false})$ $= \frac{9}{10} \times \left(1 - \left(\frac{1}{9}\right)^2 - \left(\frac{8}{9}\right)^2\right) + \frac{1}{10} \times 0 = 0.18$
$x_2 = 0$: TRUE or FALSE	{1, 3, 4, 5, 6, 7, 8, 9, 10}, {2} $\text{gini}(S) = \frac{9}{10}\text{gini}(D_{true}) + \frac{1}{10}\text{gini}(D_{false})$ $= \frac{9}{10} \times \left(1 - \left(\frac{1}{9}\right)^2 - \left(\frac{8}{9}\right)^2\right) + \frac{1}{10} \times 0 = 0.18$
$x_3 = 0$: TRUE or FALSE	{1, 2, 4, 5, 6, 7, 8, 9, 10}, {3} $\text{gini}(S) = \frac{9}{10}\text{gini}(D_{true}) + \frac{1}{10}\text{gini}(D_{false})$ $= \frac{9}{10} \times \left(1 - \left(\frac{1}{9}\right)^2 - \left(\frac{8}{9}\right)^2\right) + \frac{1}{10} \times 0 = 0.18$
$x_4 = 0$: TRUE or FALSE	{1, 5, 6, 7, 8, 9, 10}, {2, 3, 4} $\text{gini}(S) = \frac{7}{10}\text{gini}(D_{true}) + \frac{3}{10}\text{gini}(D_{false})$ $= \frac{7}{10} \times \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) + \frac{3}{10} \times 0 = 0.17$
$x_5 = 0$: TRUE or FALSE	{2, 3, 4, 6, 7, 8, 9, 10}, {1, 5} $\text{gini}(S) = \frac{8}{10}\text{gini}(D_{true}) + \frac{2}{10}\text{gini}(D_{false})$ $= \frac{8}{10} \times \left(1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2\right) + \frac{2}{10} \times 0 = 0.175$

يتبع

تابع الجدول (٤-٣) الانفصال الثنائي لعقدة الجذر والعملية الحسابية لقيمة مؤشر جيني لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع

شرط الانفصال أو الانقسام **Split Criterion**	المجموعات الفرعية الناتجة ومتوسط قيمة مؤشر جيني للانفصال **Resulting Subsets and Average Gini-Index Value of Split**
$x_6 = 0$: TRUE or FALSE	{1, 2, 4, 5, 7, 8, 9, 10}, {3, 6} $\text{gini}(S) = \frac{8}{10}\text{gini}(D_{true}) + \frac{2}{10}\text{gini}(D_{false})$ $= \frac{8}{10} \times \left(1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2\right) + \frac{2}{10} \times 0 = 0.175$
$x_7 = 0$: TRUE or FALSE	{2, 4, 8, 9, 10}, {1, 3, 5, 6, 7} $\text{gini}(S) = \frac{5}{10}\text{gini}(D_{true}) + \frac{5}{10}\text{gini}(D_{false})$ $= \frac{5}{10} \times \left(1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2\right) + \frac{5}{10} \times 0 = 0.16$
$x_8 = 0$: TRUE or FALSE	{1, 5, 6, 7, 9, 10}, {2, 3, 4, 8} $\text{gini}(S) = \frac{6}{10}\text{gini}(D_{true}) + \frac{4}{10}\text{gini}(D_{false})$ $= \frac{6}{10} \times \left(1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2\right) + \frac{4}{10} \times 0 = 0.167$
$x_9 = 0$: TRUE or FALSE	{2, 3, 4, 6, 7, 8, 10}, {1, 5, 9} $\text{gini}(S) = \frac{7}{10}\text{gini}(D_{true}) + \frac{3}{10}\text{gini}(D_{false})$ $= \frac{7}{10} \times \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) + \frac{3}{10} \times 0 = 0.17$

المثال (٤-١):

قم ببناء شجرة قرار ثنائية لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع في الجدول ٤-١.

علينا أولاً استخدام مقياس عشوائية المعلومات *(information entropy)* كمقياس لتجانس البيانات. وكما هو مبين في الشكل ٤-١، يتم تقسيم مجموعة البيانات في عقدة الجذر إلى مجموعتين فرعيتين، {2، 4، 8، 9، 10}، و{1، 3، 5، 6، 7}، والتي تظهر بالفعل متجانسة مع القيمة الهدف، $y = 1$، وليست بحاجة إلى الانفصال. بالنسبة للمجموعة الفرعية، $D=\{2,4,8,9,10\}$.

$$\text{entropy}(D) = \sum_{i=1}^{2} -P_i \log_2 P_i = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.72.$$

فيما عدا متغير الخاصية x_7 ، والذي تمّ استخدامه لتقسيم عقدة الجذر، فإن متغيرات الخاصية الثمانية الأخرى، x_1 ، x_2 ، x_3 ، x_4 ، x_5 ، x_6 ، x_8 ، x_9 يمكن استخدامها لتقسيم D.

الجدول (٤-٤) الانقسام الثنائي للعقدة الداخلية مع $D=\{2,4,5,9,10\}$ وحساب مقياس عشوائية المعلومات لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع

شرط الانفصال أو الانقسام Split Criterion	المجموعات الفرعية الناتجة ومتوسط مقياس عشوائية المعلومات للانفصال Resulting Subsets and Average Information Entropy of Split
$x_2 = 0$: TRUE or FALSE	$\{4, 8, 9, 10\}, \{2\}$ $\text{entropy}(S) = \frac{4}{5}\text{entropy}(D_{true}) + \frac{1}{5}\text{entropy}(D_{false})$ $= \frac{4}{5} \times \left(-\frac{3}{4}\log_2\frac{8}{9} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{1}{5} \times 0 = 0.64$
$x_4 = 0$: TRUE or FALSE	$\{8, 9, 10\}, \{2, 4\}$ $\text{entropy}(S) = \frac{3}{5}\text{entropy}(D_{true}) + \frac{2}{5}\text{entropy}(D_{false})$ $= \frac{3}{5} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{5} \times 0 = 0.55$
$x_8 = 0$: TRUE or FALSE	$\{9, 10\}, \{2, 4, 8\}$ $\text{entropy}(S) = \frac{2}{5}\text{entropy}(D_{true}) + \frac{3}{5}\text{entropy}(D_{false})$ $= \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{5} \times 0 = 0.4$
$x_9 = 0$: TRUE or FALSE	$\{2, 4, 8, 10\}, \{9\}$ $\text{entropy}(S) = \frac{4}{5}\text{entropy}(D_{true}) + \frac{1}{5}\text{entropy}(D_{false})$ $= \frac{4}{5} \times \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{1}{5} \times 0 = 0.64$

معايير الانفصال التي تستخدم $x_1=0$ $x_3=0$ $x_5=0$ و $x_6=0$ لا ينتج عنها تقسيم لـ D. ويوضح الجدول ٤-٤ العمليات الحسابية لمقياس عشوائية المعلومات لغرض الانفصال باستخدام x_2 x_4، x_7 x_8، x_9 وبما أن معيار الانفصال، $x_8 = 0$: *(TRUE)* أو *(FALSE)*، ينتج عنه أصغر قيمة لمتوسط مقياس عشوائية المعلومات، فإنه يتم اختيار معيار الانفصال

هذا لتقسيم $D=\{2,4,8,9,10\}$ إلى $\{9,10\}$ و$\{2,4,8\}$، والتي تبدو بالفعل متجانسة مع القيم الهدف، $y = 1$ ، وليست بحاجة إلى الانفصال. ويبين الشكل ٤-١ هذا الانفصال.

بالنسبة للمجموعة الفرعية، $D=\{9,10\}$.

$$\text{entropy}(D) = \sum_{i=1}^{2} -P_i \log_2 P_i = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1.$$

فيما عدا متغيرا الخاصية x_7 وx_8 واللذين تمّ استخدامهما لتقسيم عقدة الجذر، فإن متغيرات الخاصية السبعة الأخرى، x_1 ، x_2 ، x_3 ، x_4 ، x_5 ، x_6 ، x_9 يمكن استخدامها لتقسيم D. معايير الانفصال التي تستخدم $x_1=0$ ، $x_2=0$ $x_3=0$ $x_4=0$ $x_5=0$ و $x_6=0$ لا ينتج عنها تقسيم لـ D. معيار الانفصال $x_9 = 0$: $(TRUE)$ أو $(FALSE)$، ينتج عنه مجموعتين فرعيتين، $\{9\}$ بالقيمة الهدف $y = 1$ ، و$\{10\}$ بالقيمة الهدف $y = 1$ ، والتي تظهر متجانسةً، وليست بحاجة إلى الانفصال.

يبين الشكل ٤-١ هذا الانفصال. ولأن جميع عقد الورقة لشجرة القرار أصبحت متجانسة، فإنه يتم إيقاف عملية بناء شجرة القرار بظهور شجرة قرار كاملة كما هو مبين في الشكل ٤-١.

الجدول (٤-٥) الانقسام الثنائي للعقدة الداخلية المحتوية على $D=\{2,4,5,9,10\}$، وحساب مؤشر جيني لمجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع

شرط الانفصال أو الانقسام - Split Criterion	المجموعات الفرعية الناتجة ومتوسط قيمة مؤشر جيني للانفصال Resulting Subsets and Average Gini-Index Value of Split
$x_2 = 0$: TRUE or FALSE	{4, 8, 9, 10}, {2} $\text{gini}(S)=\frac{4}{5}\text{gini}(D_{true})+\frac{1}{5}\text{gini}(D_{false})$ $=\frac{4}{5}\times\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)+\frac{1}{5}\times 0=0.3$
$x_4 = 0$: TRUE or FALSE	{8, 9, 10}, {2, 4} $\text{gini}(S)=\frac{3}{5}\text{gini}(D_{true})+\frac{2}{5}\text{gini}(D_{false})$ $=\frac{3}{5}\times\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)+\frac{2}{5}\times 0=0.27$
$x_8 = 0$: TRUE or FALSE	{9, 10}, {2, 4, 8} $\text{gini}(S)=\frac{2}{5}\text{gini}(D_{true})+\frac{3}{5}\text{gini}(D_{false})$ $=\frac{2}{5}\times\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right)+\frac{3}{5}\times 0=0.2$
$x_9 = 0$: TRUE or FALSE	{2, 4, 8, 10}, {9} $\text{gini}(S)=\frac{4}{5}\text{gini}(D_{true})+\frac{1}{5}\text{gini}(D_{false})$ $=\frac{4}{5}\times\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)+\frac{1}{5}\times 0=0.3$

وسوف نوضح الآن عملية بناء شجرة قرار باستخدام مؤشر جيني كمقياس لتجانس البيانات. كما هو موضح سابقاً، يتم تقسيم مجموعة البيانات في عقدة الجذر إلى مجموعتين فرعيتين،

{2، 4، 8، 9، 10}، و{1، 3، 5، 6، 7}، والتي تظهر بالفعل متجانسة مع القيمة الهدف،حيث $y = 1$ ، وليست بحاجة إلى الانقسام. بالنسبة للمجموعة الفرعية، $D=\{2,4,8,9,10\}$،

$$\text{gini}(D) = 1 - \sum_{i=1}^{c} P_i^2 = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.32.$$

معيار الانفصال باستخدام أيٍّ من $x_1=0$ ، $x_3=0$ $x_5=0$ و $x_6=0$ لا ينتج عنه انقسام لـ D. يوضح الجدول ٤-٥ عملية حساب قيم مؤشر جيني للانفصالات باستخدام x_2 ، x_4 ، x_7 ، x_8 ، x_9 . بما أن معيار الانفصال، $x_8=0$: (TRUE) أو (FALSE) ، ينتج عنه أصغر قيمة لمتوسط مؤشر جيني للانفصال، يتم اختيار معيار الانفصال هذا لتقسيم $D=\{2,4,8,9,10\}$ إلى {9، 10} و{2,4,8}، والتي تظهر فعلياً متجانسة مع القيم الهدف، $y = 1$ ، وهي ليست بحاجة إلى الانفصال.

بالنسبة للمجموعة الفرعية، $D=\{9,10\}$

$$\text{gini}(D) = 1 - \sum_{i=1}^{c} P_i^2 = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5.$$

فيما عدا متغيرا الخاصية $x7$ $x8$ ، واللذين تمّ استخدامها لتقسيم عقدة الجذر، فإن متغيرات الخاصية السبعة الأخرى، $x1$ $x2$ $x3$ ، $x4$ ، $x5$ $x6$ و$x9$ يمكن استخدامها لتقسيم D. معايير الانفصال التي تستخدم $x_1=0$ ، $x_2=0$ $x_3=0$ $x_4=0$ $x_5=0$ و $x_6=0$ لا ينتج عنها تقسيم لـ D. معيار الانفصال $x_8=0$: (TRUE) أو (FALSE)، ينتج عنه مجموعتان فرعيتان، {9} بقيمة الهدف $y = 1$ ، و{10} بقيمة الهدف $y=0$، والتي تبدو متجانسة، وليست بحاجة إلى الانفصال. ولأن جميع عقد الورقة لشجرة القرار أصبحت متجانسة، فإنه يتم إيقاف عملية بناء شجرة القرار بظهور شجرة قرار كاملة، وهي شجرة القرار نفسه التي تستخدم مقياس عشوائية المعلومات كمقياس لتجانس البيانات.

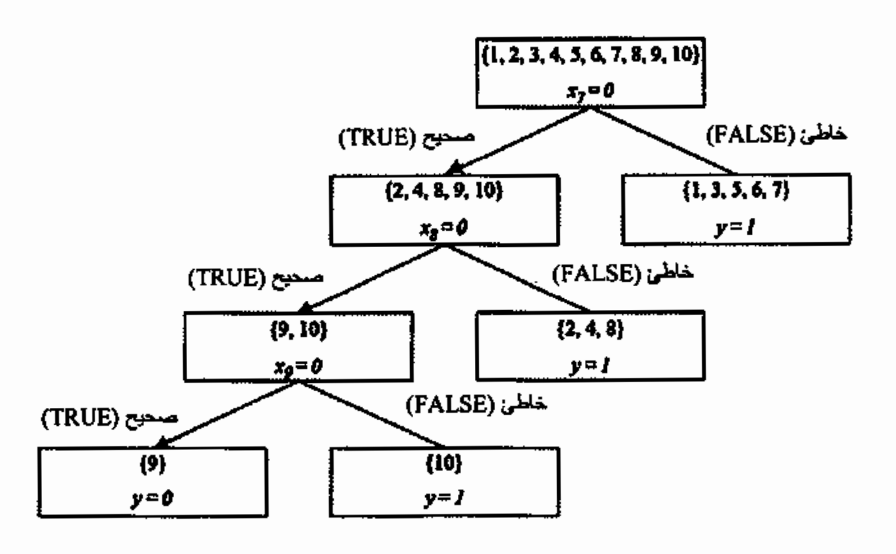
٤-١-٥ تصنيف البيانات باستخدام شجرة القرار
(Classifying Data Using a Decision Tree):

يتم استخدام شجرة القرار لتصنيف سجل البيانات عن طريق تمرير سجل البيانات إلى عقدة الورقة في شجرة القرار باستخدام قيم متغيرات الخاصية، وإسناد قيمة الهدف الخاصة بعقدة الورقة لسجل البيانات.

يبرز الشكل ٤-٣ مسار تمرير سجل بيانات التدريب باللون الداكن، للسجل رقم ١٠ في الجدول ٤-١، ابتداءً من عقدة الجذر إلى عقدة الورقة بقيمة لمتغير الهدف، $y=0$. ومن ثم، فإنه يتم تصنيف سجل البيانات رقم ١٠ بدون عطل في النظام. بالنسبة لسجلات البيانات في مجموعة البيانات الاختبارية الخاصة بالكشف عن الأعطال بنظام التصنيع الموضحة في الجدول ٤-٦، فإنه يتم الحصول على القيم الهدف الخاصة بالسجلات باستخدام شجرة القرار في الشكل ٤-١، وهي موضحة في الجدول ٤-٦. يسلط الشكل ٤-٤ الضوء على مسار تمرير سجل بيانات اختباري للسجل رقم ١ في الجدول ٤-٦ من عقدة الجذر إلى عقدة الورقة ذات القيمة الهدف، $y = 1$. ومن ثم، يتم تصنيف سجل البيانات هذا على أنه يحتوي على عطل في النظام.

الشكل (٤-٣)

تصنيف سجل بيانات بدون عطل نظام باستخدام شجرة القرار الخاصة بالكشف عن أعطال نظام التصنيع



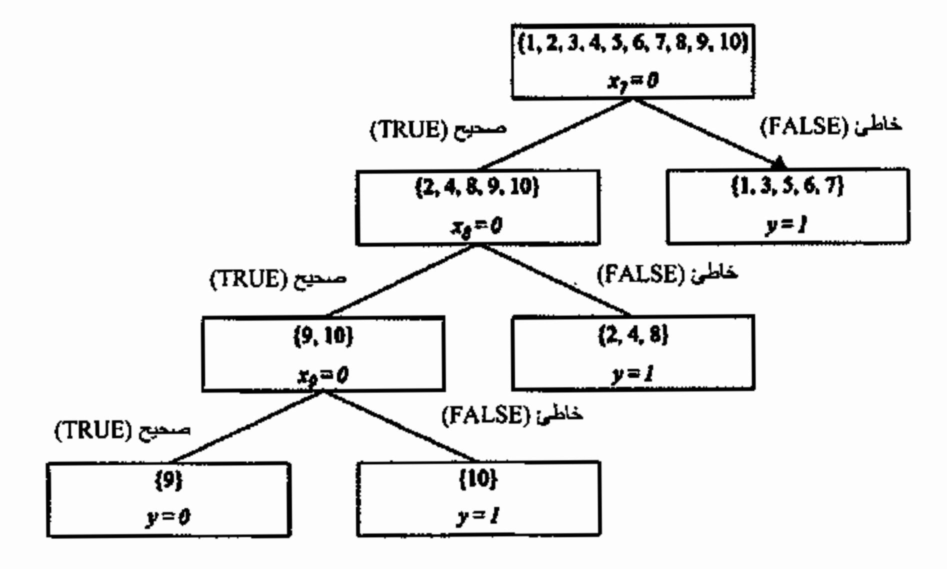
الجدول (٤-٦)

تصنيف سجلات البيانات لمجموعة البيانات الاختبارية الخاصة بالكشف عن أعطال نظام التصنيع

متغير الهدف y - Target Variable (أعطال النظام - System Faults)		Attribute Variables - متغيرات الخاصية (Quality of Parts - جودة وحدات المنتج)									رقم الحالة Instance (الآلة المعطلة - Faulty Machine)
القيمة المصنفة (Classified Value)	القيمة الفعلية (True Value)	x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	
1	1	1	1	1	0	1	1	0	1	1	1 (M1,M2)
1	1	0	1	1	1	0	1	1	1	0	2(M2,M3)
1	1	1	0	1	1	1	0	1	0	1	3(M1,M3)
1	1	1	1	1	0	1	1	0	0	1	4(M1,M4)
1	1	1	0	1	1	1	0	0	0	1	5(M1,M6)
1	1	0	1	1	1	0	1	0	1	0	6(M2,M6)
1	1	0	1	1	0	1	1	0	1	0	7(M2,M5)
1	1	1	0	1	1	1	0	1	0	0	8(M3,M5)
1	1	0	1	1	0	0	1	0	0	0	9(M4,M7)
1	1	0	1	1	0	1	0	0	0	0	10(M5,M8)
1	1	1	1	1	1	0	1	1	0	0	11(M3,M9)
1	1	1	1	1	0	1	0	0	0	1	12(M1,M8)
1	1	1	1	1	1	1	1	1	1	1	13(M1,M2,M3)
1	1	1	1	1	1	1	1	1	1	0	14(M2,M3,M5)
1	1	1	1	1	1	0	1	1	1	0	15(M2,M3,M9)
1	1	1	1	1	1	1	0	0	0	1	16(M1,M6,M8)

الشكل (٤-٤)

تصنيف سجل بيانات لأعطال متعددة الآلات باستخدام شجرة قرار خاصة بالكشف عن أعطال نظام التصنيع



٤-٢ تعلُّم شجرة القرار غير الثنائية (Learning a Nonbinary Decision Tree):

يوجد ثلاث قيم نوعية لمتغير الخاصية، العمر *(age)*، في مجموعة البيانات الخاصة بالعدسات في الجدول ١-٣، والقيم هي: شاب *(Young)*، ما قبل الشيخوخة، *(Pre-presbyopic)*، والشيخوخة (*Presbyopic*). إذا أردنا بناء شجرة قرار ثنائية لمجموعة البيانات هذه، فنحن بحاجة إلى تحويل القيم النوعية الثلاثة لمتغير الخاصية العمر *(age)* إلى قيمتين نوعيتين عند استخدام العمر لتقسيم عقدة الجذر. قد نضع الفئتين: شاب وما قبل الشيخوخة معاً في فئة واحدة، وتكون الفئة: الشيخوخة في فئة أخرى، ويكون معيار الانفصال كما يلي: العمر = الشيخوخة صحيح أو خطأ. بإمكاننا أيضاً وضع الفئة: شاب كفئة واحده والفئتين: ما قبل الشيخوخة، والشيخوخة معاً في فئة أخرى، ويكون شرط أو معيار الانفصال كما يلي: العمر = شاب: صحيح أو خطأ. لكن، يمكننا بناء شجرة قرار غير ثنائية للسماح بتقسيم مجموعة بيانات لعقدة ما إلى أكثر من مجموعتين فرعيتين باستخدام القيم النوعية المتعددة لكل فرع من الانقسام.

المثال ٤-٢ يوضح كيفية بناء شجرة قرار غير ثنائية لمجموعة بيانات العدسات.

المثال ٤-٢:

قم ببناء شجرة قرار غير ثنائية لمجموعة بيانات العدسات في الجدول ١-٣. إذا استُخدم متغير الخاصية، العمر - *age*، لـتقسيم عقدة الجذر لمجموعة بيانات العدسات، فإنه يمكن استخدام كل القيم النوعية الثلاثة لـ "العمر" لتقسيم مجموعة سجلات البيانات المكونة من ٢٤ سجل في عقدة الجذر باستخدام معيار الانقسام، العمر = شاب، قبل الشيخوخة، أو الشيخوخة، كما هو موضح في الشكل ٤-٥. يتم استخدام مجموعة البيانات المكونة من ٢٤ سجل موضحة في الجدول ١-٣ على أنها مجموعة البيانات التدريبية، D، في عقدة الجذر لشجرة القرار غير الثنائية. في مجموعة بيانات العدسات، المتغير الهدف له ثلاث قيم نوعية، وهي العدسات غير اللاصقة الخارجية (*Non-Contact*) موجودة في ١٥ سجل، والعدسات اللاصقة الطرية (*Soft-Contact*) موجودة في ٥ سجلات، والعدسات اللاصقة الصلبة (*Hard-Contact*) موجودة في ٤ سجلات. باستخدام مقياس عشوائية المعلومات كمقياس لتجانس البيانات، يصبح لدينا:

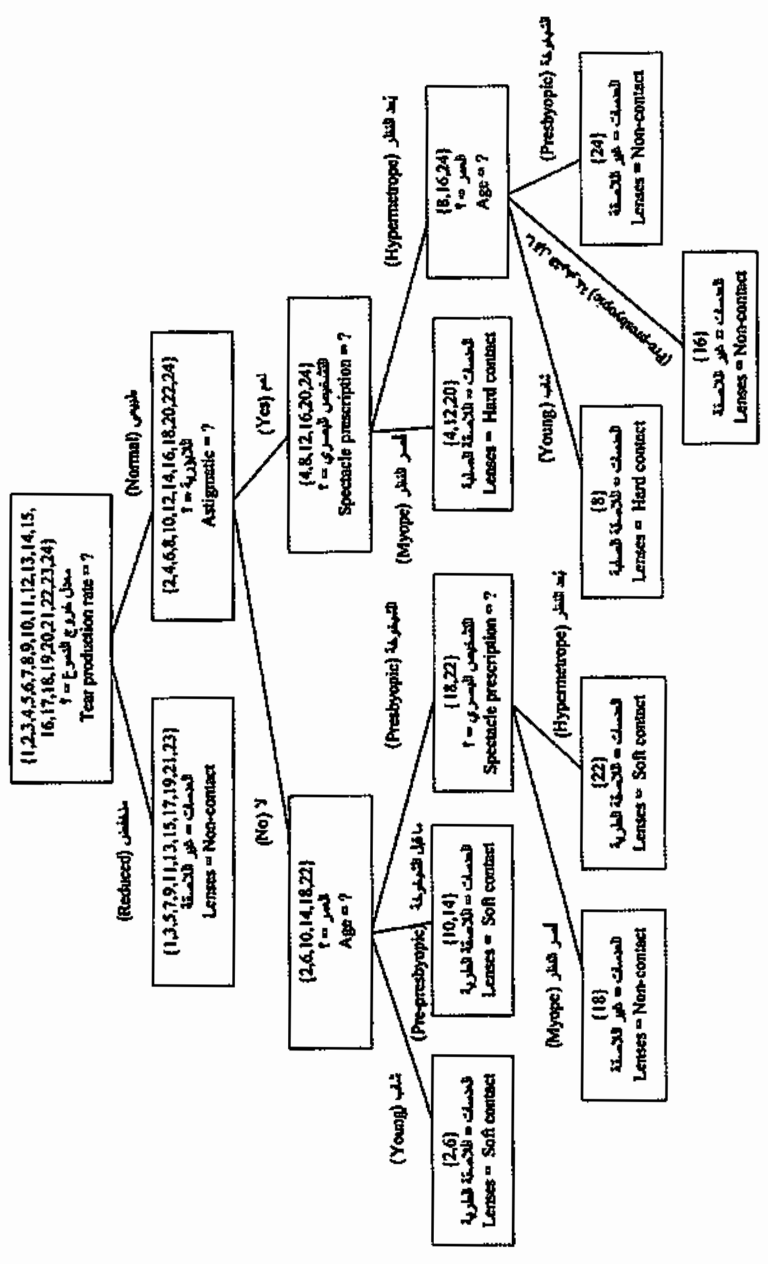
$$\begin{aligned} \text{entropy}(D) &= \sum_{i=1}^{3} -P_i \log_2 P_i \\ &= -\frac{15}{24}\log_2\frac{15}{24} - \frac{5}{24}\log_2\frac{5}{24} - \frac{4}{24}\log_2\frac{4}{24} \\ &= 1.3261. \end{aligned}$$

ويبين الجدول ٤-٧ عملية حساب مقياس عشوائية المعلومات لتقسيم فرعية عقدة الجذر باستخدام معيار الانفصال، معدل خروج الدموع *(tear production rate)* = منخفض *(reduced)* أو عادي *(normal)*، والذي ينتج عنه مجموعة فرعية متجانسة وأرقام سجلاتها {1، 3، 5، 7، 9، 11، 13، 15، 17، 19، 21، 23} ومجموعة فرعية أخرى غير متجانسة *{2,4,6,8,10,12,14,16,18,20,22,24}*. ويبين الجدول ٤-٨ عملية حساب مؤشر مقياس عشوائية لتقسيم العقدة المحتوية على مجموعة البيانات {2، 4، 6، 8، 10، 12، 14، 16، 18، 20، 22، 24} باستخدام معيار الانقسام، اللابؤرية

(astigmatic) = لا *(No)* أو نعم *(Yes)*، والتي تنتج عنها مجموعتان فرعيتان {2، 6، 10، 14، 18، 22} و{4، 8، 12، 16، 20، 24}.

ويبين الجدول ٤-٩ عملية حساب مقياس عشوائية المعلومات لتقسيم العقدة المحتوية على مجموعة البيانات {2، 6، 10، 14، 18، 22} باستخدام معيار الانقسام، العمر *(Age)* = شاب *(Young)*، قبل الشيخوخة *(Pre-presbyopic)* ، أو الشيخوخة *(Presbyopic)*، التي تنتج ثلاثة مجموعات فرعية {2، 6}، {10، 14}، و{18، 22}. يتم تقسيم هذه المجموعات الفرعية علاوةً على ذلك باستخدام معيار الانقسام، الوصفة الطبية *(spectacle prescription)* = قصر النظر *(myope)* أو بُعد النظر *(hypermetrope)*، للحصول على عُقد الورقة ذات مجموعات بيانات متجانسة. ويبين الجدول ٤-١٠ عملية حساب مقياس عشوائية المعلومات لتقسيم العقدة المحتوية على مجموعة البيانات {4، 8، 12، 16، 20، 24} باستخدام معيار الانقسام، الوصفة الطبية = قصر النظر أو بُعد النظر، والتي تنتج مجموعتين فرعيتين {4، 12، 20} و{8، 16، 24}. ويتم تقسيم هذه المجموعات الفرعية باستخدام معيار الانقسام، العمر = شاب، قبل الشيخوخة، أو الشيخوخة، لإنتاج عقد الورقة ذات مجموعات بيانات متجانسة. ويبين الشكل ٤-٥ شجرة القرار غير الثنائية الكاملة لمجموعة بيانات العدسات.

الشكل (٤-٥) شجرة القرار لمجموعة بيانات العدسات



الجدول (٤-٧) الانفصال غير الثنائي لعقدة الجذر وعملية حساب مقياس عشوائية المعلومات لمجموعة بيانات العدسات

شرط الانفصال أو الانقسام **Split Criterion**	المجموعات الفرعية الناتجة ومتوسط مقياس عشوائية المعلومات للانفصال **Resulting Subsets and Average Information Entropy of Split**
Age = Young, Pre-presbyopic, or Presbyopic العمر = شاب، ما قبل الشيخوخة، أو الشيخوخة	{1, 2, 3, 4, 5, 6, 7, 8}, {9, 10, 11, 12, 13, 14, 15, 16}, {17, 18, 19, 20, 21, 22, 23, 24} $\text{entropy}(S)=\frac{8}{24}\text{entropy}(D_{Young})+\frac{8}{24}\text{entropy}(D_{Pre\text{-}presbyopic})$ $+\frac{8}{24}\text{entropy}(D_{Presbyopic})$ $=\frac{8}{24}\times\left(-\frac{4}{8}\log_2\frac{4}{8}-\frac{2}{8}\log_2\frac{2}{8}-\frac{2}{8}\log_2\frac{2}{8}\right)$ $+\frac{8}{24}\times\left(-\frac{5}{8}\log_2\frac{5}{8}-\frac{2}{8}\log_2\frac{2}{8}-\frac{1}{8}\log_2\frac{1}{8}\right)$ $+\frac{8}{24}\times\left(-\frac{6}{8}\log_2\frac{6}{8}-\frac{1}{8}\log_2\frac{1}{8}-\frac{1}{8}\log_2\frac{1}{8}\right)=1.2867$
Spectacle Prescription = Myope or Hypermetrope التشخيص البصري = قُصر النظر أو بُعد النظر	{1, 2, 3, 4, 9, 10, 11, 12, 17, 18, 19, 20}, {5, 6, 7, 8, 13, 14, 15, 16, 21, 22, 23, 24} $\text{entropy}(S)=\frac{12}{24}\text{entropy}(D_{Myope})+\frac{12}{24}\text{entropy}(D_{Hypermetrope})$ $=\frac{12}{24}\times\left(-\frac{7}{12}\log_2\frac{7}{12}-\frac{2}{12}\log_2\frac{2}{12}-\frac{3}{12}\log_2\frac{3}{12}\right)$ $+\frac{12}{24}\times\left(-\frac{8}{12}\log_2\frac{8}{12}-\frac{3}{12}\log_2\frac{3}{12}-\frac{1}{12}\log_2\frac{1}{12}\right)$ $=1.2866$
Astigmatic = No or Yes اللابؤرية = لا أو نعم	{1, 2, 5, 6, 9, 10, 13, 14, 17, 18, 21, 22}, {3, 4, 7, 8, 11, 12, 15, 16, 19, 20, 23, 24} $\text{entropy}(S)=\frac{12}{24}\text{entropy}(D_{No})+\frac{12}{24}\text{entropy}(D_{Yes})$ $=\frac{12}{24}\times\left(-\frac{7}{12}\log_2\frac{7}{12}-\frac{5}{12}\log_2\frac{5}{12}-\frac{0}{12}\log_2\frac{0}{12}\right)$ $+\frac{12}{24}\times\left(-\frac{8}{12}\log_2\frac{8}{12}-\frac{4}{12}\log_2\frac{4}{12}-\frac{0}{12}\log_2\frac{0}{12}\right)$ $=0.9491$
Tear Production Rate = Reduced or Normal معدل خروج الدموع = منخفض أو طبيعي	{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23}, {2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24} $\text{entropy}(S)=\frac{12}{24}\text{entropy}(D_{Reduced})+\frac{12}{24}\text{entropy}(D_{Normal})$ $=\frac{12}{24}\times\left(-\frac{12}{12}\log_2\frac{12}{12}-\frac{0}{12}\log_2\frac{0}{12}-\frac{0}{12}\log_2\frac{0}{12}\right)$ $+\frac{12}{24}\times\left(-\frac{3}{12}\log_2\frac{3}{12}-\frac{5}{12}\log_2\frac{5}{12}-\frac{4}{12}\log_2\frac{4}{12}\right)$ $=0.7773$

الجدول (٤-٨) الانفصال غير الثنائي للعقدة الداخلية {2، 4، 6، 8، 10، 12، 14، 16، 18، 20، 22، 24}، وعملية حساب مقياس عشوائية المعلومات لمجموعة بيانات العدسات

شرط الانفصال أو الانقسام **Split Criterion**	المجموعات الفرعية الناتجة ومتوسط مقياس عشوائية المعلومات للانفصال **Resulting Subsets and Average Information Entropy of Split**
Age = Young, Pre-presbyopic, or Presbyopic العمر = شاب، ما قبل الشيخوخة، أو الشيخوخة	{2, 4, 6, 8}, {10, 12, 14, 16}, {18, 20, 22, 24} $\text{entropy}(S)=\frac{4}{12}\text{entropy}(D_{Young})+\frac{4}{12}\text{entropy}(D_{Pre-presbyopic})+\frac{4}{12}\text{entropy}(D_{Presbyopic})$ $=\frac{4}{12}\times\left(-\frac{0}{4}\log_2\frac{0}{4}-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}\right)+\frac{4}{12}\times\left(-\frac{1}{4}\log_2\frac{1}{4}-\frac{2}{4}\log_2\frac{2}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)+\frac{4}{12}\times\left(-\frac{2}{4}\log_2\frac{2}{4}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)$ $=1.3333$
Spectacle Prescription = Myope or Hypermetrope التشخيص البصري = قُصر النظر أو بُعد النظر	{2, 4, 10, 12, 18, 20}, {6, 7, 14, 16, 22, 24} $\text{entropy}(S)=\frac{6}{12}\text{entropy}(D_{Myope})+\frac{6}{12}\text{entropy}(D_{Hypermetrope})$ $=\frac{6}{12}\times\left(-\frac{1}{6}\log_2\frac{1}{6}-\frac{2}{6}\log_2\frac{2}{6}-\frac{3}{6}\log_2\frac{3}{6}\right)+\frac{6}{12}\times\left(-\frac{2}{6}\log_2\frac{2}{6}-\frac{3}{6}\log_2\frac{3}{6}-\frac{1}{6}\log_2\frac{1}{6}\right)$ $=1.4591$
Astigmatic = No or Yes اللابؤرية = لا أو نعم	{2, 6, 10, 14, 18, 22}, {4, 8, 12, 16, 20, 24} $\text{entropy}(S)=\frac{6}{12}\text{entropy}(D_{No})+\frac{6}{12}\text{entropy}(D_{Yes})$ $=\frac{6}{12}\times\left(-\frac{1}{6}\log_2\frac{1}{6}-\frac{5}{6}\log_2\frac{5}{6}-\frac{0}{6}\log_2\frac{0}{6}\right)+\frac{6}{12}\times\left(-\frac{2}{6}\log_2\frac{2}{6}-\frac{0}{6}\log_2\frac{0}{6}-\frac{4}{6}\log_2\frac{4}{6}\right)$ $=0.7842$

الجدول (٤-٩)

الانفصال غير الثنائي للعقدة الداخلية {2، 6، 10، 14، 18، 22}، وعملية حساب مقياس عشوائية المعلومات لمجموعة بيانات العدسات.

شرط الانفصال أو الانقسام Split Criterion	المجموعات الفرعية الناتجة ومتوسط مقياس عشوائية المعلومات للانفصال Resulting Subsets and Average Information Entropy of Split
Age = Young, Pre-presbyopic, or Presbyopic العمر = شاب، ما قبل الشيخوخة، أو الشيخوخة	{2, 6}, {10, 14}, {18, 22} $\text{entropy}(S)=\frac{2}{6}\text{entropy}(D_{Young})+\frac{2}{6}\text{entropy}(D_{Pre\text{-}presbyopic})+\frac{2}{6}\text{entropy}(D_{Presbyopic})$ $=\frac{2}{6}\times\left(-\frac{0}{2}\log_2\frac{0}{2}-\frac{2}{2}\log_2\frac{2}{2}-\frac{0}{2}\log_2\frac{0}{2}\right)+\frac{2}{6}\times\left(-\frac{0}{2}\log_2\frac{0}{2}-\frac{2}{2}\log_2\frac{2}{2}-\frac{0}{2}\log_2\frac{0}{2}\right)+\frac{2}{6}\times\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}-\frac{0}{2}\log_2\frac{0}{2}\right)$ $=0.3333$
Spectacle Prescription = Myope or Hypermetrope التشخيص البصري = قُصر النظر أو بُعد النظر	{2, 10, 18}, {6, 14, 22} $\text{entropy}(S)=\frac{3}{6}\text{entropy}(D_{Myope})+\frac{3}{6}\text{entropy}(D_{Hypermetrope})$ $=\frac{3}{6}\times\left(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}-\frac{0}{3}\log_2\frac{0}{3}\right)+\frac{3}{6}\times\left(-\frac{0}{3}\log_2\frac{0}{3}-\frac{3}{3}\log_2\frac{3}{3}-\frac{0}{3}\log_2\frac{0}{3}\right)$ $=0.4591$

٤-٣ التعامل مع القيم الرقمية والقيم المفقودة لمتغيرات الخاصية (Handling Numeric and Missing Values of Attribute Variables):

إذا كانت مجموعة البيانات تحتوي على متغير خاصية رقمي، يحتاج المتغير إلى أن يتحول إلى متغير نوعي قبل استخدامه لغرض بناء شجرة القرار. سنستعرض الطريقة الشائعة لعمل هذا التحول. لنفترض أن لدينا متغير خاصية رقمي، x لديه القيم الرقمية التالية في مجموعة

البيانات التدريبية، a_k ...... a_1 ،a_2 ، والتي يتم فرزها بترتيب متزايد تصاعدي. النقطة أو القيمة الوسطى لقيمتين رقميتين متجاورتين، a_i و a_j، يتم حسابها على النحو التالي:

$$c_i = \frac{a_i + a_j}{2} \qquad (٧-٤)$$

الجدول (٤-١٠) الانفصال غير الثنائي للعقدة الداخلية {4، 8، 12، 16، 20، 24} وعملية حساب مقياس عشوائية المعلومات لمجموعة بيانات العدسات.

شرط الانفصال أو الانقسام Split Criterion	المجموعات الفرعية الناتجة ومتوسط مقياس عشوائية المعلومات للانفصال Resulting Subsets and Average Information Entropy of Split
Age = Young, Pre-presbyopic, or Presbyopic العمر = شاب، ما قبل الشيخوخة، أو الشيخوخة	{4, 8}, {12, 16}, {20, 24} $\text{entropy}(S)=\frac{2}{6}\text{entropy}(D_{Young})+\frac{2}{6}\text{entropy}(D_{Pre-presbyopic})+\frac{2}{6}\text{entropy}(D_{Presbyopic})$ $=\frac{2}{6}\times\left(-\frac{0}{2}\log_2\frac{0}{2}-\frac{0}{2}\log_2\frac{0}{2}-\frac{2}{2}\log_2\frac{2}{2}\right)$ $+\frac{2}{6}\times\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{0}{2}\log_2\frac{0}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)$ $+\frac{2}{6}\times\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{0}{2}\log_2\frac{0}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)$ $=0.6667$
Spectacle Prescription = Myope or Hypermetrope التشخيص البصري = قصر النظر أو بُعد النظر	{4, 12, 20}, {8, 16, 24} $\text{entropy}(S)=\frac{3}{6}\text{entropy}(D_{Myope})+\frac{3}{6}\text{entropy}(D_{Hypermetrope})$ $=\frac{3}{6}\times\left(-\frac{0}{3}\log_2\frac{0}{3}-\frac{0}{3}\log_2\frac{0}{3}-\frac{3}{3}\log_2\frac{3}{3}\right)$ $+\frac{3}{6}\times\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{0}{3}\log_2\frac{0}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)$ $=0.4591$

باستخدام c_i حيث، $i=1, \ldots, k-1$ يمكننا إنشاء القيم النوعية التالية والتي عددها $k+1$ كقيم لـ x :

$$
\begin{array}{ll}
\text{Category } 1: & x \leq c_1 \\
\text{Category } 2: & c_1 < x \leq c_1 \\
 & \vdots \\
\text{Category } k: & c_{k-1} < x \leq c_k \\
\text{Category } k+1: & c_k < x.
\end{array}
$$

يتم تحويل القيمة الرقمية لـ x إلى قيمة نوعية وفقاً للتعريف المذكور آنفاً للقيم النوعية. على سبيل المثال، إذا $c_1<x\leq c_2$، فإن القيمة النوعية لـ x هي الفئة (Category 2).

في العديد من مجموعات البيانات، قد نجد متغير خاصية بدون قيمة في سجل بيانات ما. على سبيل المثال، إذا كان هناك متغيرات خاصية للاسم، والعنوان، وعنوان البريد الإلكتروني للعملاء في قاعدة بيانات متجر ما، قد لا يكون هناك عنوان البريد الإلكتروني لعميل معين. وهو ما يعني، أنه قد تكون لدينا عناوين بريد إلكتروني مفقودة لبعض العملاء. إحدى الطرق لمعالجة سجل بيانات يحتوي على قيمة مفقودة هو بتجاهل سجل البيانات. لكن، عندما تكون مجموعة البيانات التدريبية صغيرة، فنحن بحاجة إلى جميع سجلات البيانات لمجموعة البيانات التدريبية حتى تتمكن من بناء شجرة القرار. ولاستخدام سجل بيانات يحتوي على قيمة مفقودة، قد نكون بحاجة إلى تقدير القيمة المفقودة، واستخدام القيمة التقديرية لملء القيمة المفقودة. بالنسبة لمتغير الخاصية النوعي، يمكن تقدير القيمة المفقودة الخاصة به لتكون القيمة الأكثر شيوعاً في غالبية سجلات البيانات في مجموعة البيانات التدريبية التي لها نفس القيمة لمتغير الهدف مثل تلك الموجودة في سجل البيانات ذو القيمة المفقودة لمتغير الخاصية. وبالنسبة لمتغير الخاصية الرقمي، يمكن تقدير القيمة المفقودة الخاصة به لتكون قيمة متوسط القيم التي يتم اتخاذها من قبل سجلات البيانات في مجموعة البيانات التدريبية التي لها قيمة المتغير الهدف نفسه مثل تلك الموجودة في سجل البيانات ذي القيمة المفقودة لمتغير الخاصية. وترد أساليب أخرى لتقدير القيمة المفقودة في (*Ye, 2003*).

٤-٤ التعامل مع متغير الهدف الرقمي وبناء شجرة الانحدار
(Handling a Numeric Target Variable and Constructing a Regression Tree):

إذا كان لدينا متغير هدف رقمي، فإنه لا يمكن تطبيق مقاييس تجانس البيانات، مثل: مقياس عشوائية المعلومات، ومؤشر جيني. ويقدم بريمان وآخرون (*Breiman et al., 1984*) المعادلة رقم ٤-٧ لحساب متوسط اختلاف القيم عن قيمة متوسطها، R، واستخدامه لقياس تجانس البيانات لبناء شجرة الانحدار عندما تكون قيم المتغير الهدف رقمية. متوسط الاختلاف للقيم في مجموعة بيانات من قيمة متوسطها يشير إلى مدى كون القيم متشابهة أو متجانسة. فكلما كانت قيمة R أصغر، كانت مجموعة البيانات أكثر تجانساً. المعادلة ٤-٩ تبين عملية حساب متوسط قيمة R بعد الانفصال:

$$R(D) = \sum_{y \in D} (y - \bar{y})^2 \qquad (٤-٨)$$

$$\bar{y} = \frac{\sum_{y \in D} y}{n} \qquad (٤-٩)$$

$$R(S) = \sum_{v \in Values(S)} \frac{|D_v|}{|D|} R(D_v) \qquad (٤-١٠)$$

مجموعة البيانات الخاصة بمكوك الفضاء في الجدول ١-٢، تحتوي متغير هدف رقمي، وأربعة متغيرات خاصية رقمية. يتم حساب قيمة R لمجموعة البيانات D لسجلات البيانات الـ ٢٣ في عقدة الجذر لشجرة الانحدار كما يلي:

$$\bar{y} = \frac{\sum_{y \in D} y}{n}$$

$$= \frac{0+1+0+0+0+0+0+0+1+1+1+0+0+2+0+0+0+0+0+0+0+0+1}{23}$$

$$= 0.3043$$

$$R(D) = \sum_{y \in D} (y - \bar{y})^2 = (0 - 0.3043)^2 + (1 - 0.3043)^2 + (0 - 0.3043)^2 + (0 - 0.3043)^2$$
$$+ (0 - 0.3043)^2 + (0 - 0.3043)^2 + (0 - 0.3043)^2 + (0 - 0.3043)^2 + (1 - 0.3043)^2$$
$$+ (1 - 0.3043)^2 + (1 - 0.3043)^2 + (0 - 0.3043)^2 + (0 - 0.3043)^2 + (2 - 0.3043)^2$$
$$+ (0 - 0.3043)^2 + (0 - 0.3043)^2 + (0 - 0.3043)^2 + (0 - 0.3043)^2 + (0 - 0.3043)^2$$
$$+ (0 - 0.3043)^2 + (0 - 0.3043)^2 + (0 - 0.3043)^2 + (1 - 0.3043)^2$$
$$= 6.8696$$

وغالباً ما يتم استخدام متوسط قيم الهدف لسجلات البيانات الموجودة في عقدة الورقة لشجرة القرار ذات متغير الهدف الرقمي، كقيمة هدف لعقدة الورقة. عند تمرير سجل بيانات على طول شجرة القرار لتحديد القيمة الهدف لسجل البيانات، يتم إسناد القيمة الهدف لعقدة الورقة حيث يصل سجل البيانات كقيمة الهدف الخاص بـ سجل البيانات. وتُسمى شجرة القرار ذات المتغير الهدف الرقمي بشجرة الانحدار (*regression tree*).

٤-٥ مزايا وعيوب خوارزمية شجرة القرار
(Advantages and Shortcomings of the Decision Tree algorithm):

إن من مميزات استخدام خوارزمية شجرة القرار لتعلُّم أنماط التصنيف والتنبؤ هو التعبير الصريح لأنماط التصنيف والتنبؤ لشجرة القرار والانحدار. تكشف شجرة القرار في الشكل ٤-١ عن ثلاثة أنماط خاصة بجودة قطع الغيار، الأمر الذي يؤدي إلى ثلاثة من عقد الورقة ذات التصنيف "عطل في النظام"، على التوالي.

- $x_7=1$
- $x_7=0$ & $x_8=1$
- $x_7=0$ & $x_8=0$ & $x_9=1$

والنمط التالي الخاص بجودة القطع لعقدة ورقة واحدة ذات تصنيف "بدون عطل بالنظام":

- $x7=0$ & $x8=0$ & $x9=0$

أنماط التصنيف الصريحة المذكورة أعلاه تكشف عن المعرفة الأساسية التالية للكشف عن أعطال نظام التصنيع هذا:

- من بين متغيرات الجودة التسعة، يتضح أن متغيرات الجودة الثلاثة، x_7 ، x_8 x_9 هي ذات أهمية للكشف عن أعطال نظام التصنيع. تسمح لنا هذه المعرفة بالحد من تكلفة فحص جودة القطع من خلال فحص جودة القطع بعد الآلات السابعة *M7*، والثامنة *M8* والتاسعة *M9* فقط بدلاً من فحص الآلات التسع كلها.
- إذا كان أحد هذه المتغيرات الثلاثة، x_7 ، x_8 x_9 يظهر فشلاً في الجودة، فإن النظام يكون به عطل؛ وخلاف ذلك، لا يوجد لدى النظام عطل.

هناك أيضاً قصور لدى شجرة القرار عند التعبير عن أنماط التصنيف والتنبؤ لأنها تستخدم متغير خاصية واحد فقط في معيار الانفصال. هذا قد يؤدي إلى شجرة قرار كبيرة. وفي شجرة القرار الكبيرة، يكون من الصعب أن نرى أنماط واضحة للتصنيف والتنبؤ. على سبيل المثال، في الفصل ١، قدمنا نمط التصنيف التالي لمجموعة بيانات البالون في الجدول ١-١:

IF (Color = Yellow AND Size = Small) OR (Age = Adult AND Act = Stretch), THEN Inflated = T; OTHERWISE, Inflated = f.

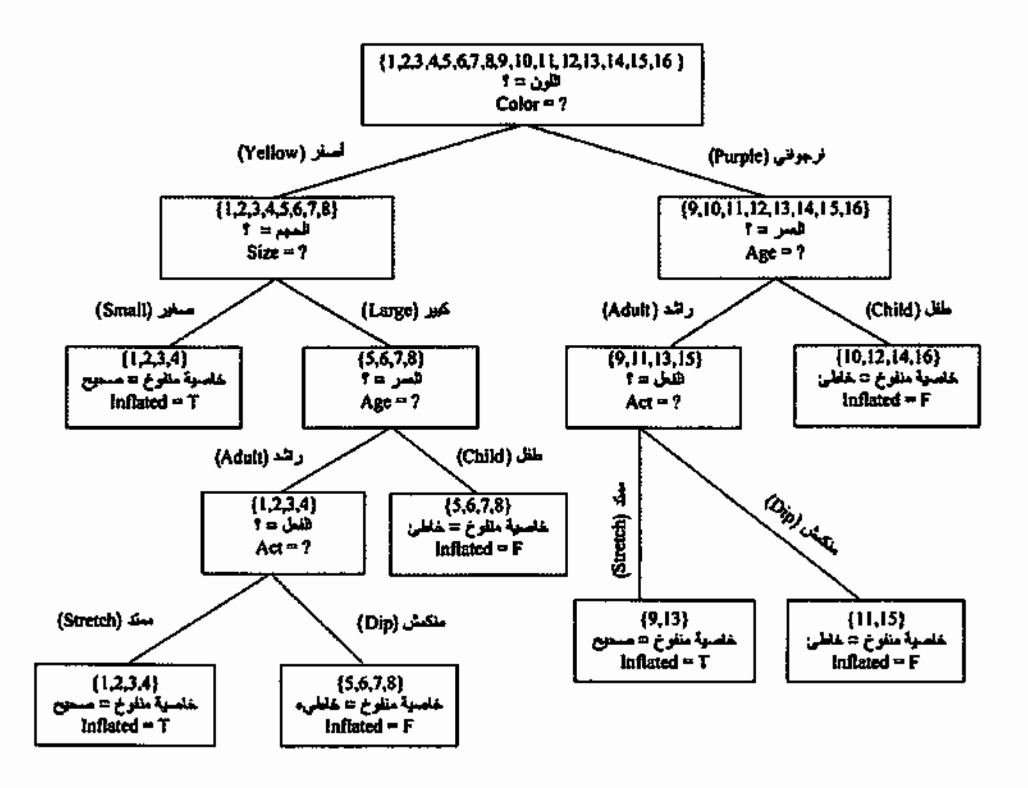
إذا كان (اللون = أصفر، والحجم = صغير) أو (العمر = راشد والفعل = ممتد)، إذن تكون خاصية منفوخ = T (أي "صحيح")؛ وإلا تكون خاصية منفوخ = F (أي "خاطيء").

هذا النمط لتصنيف قيمة الهدف لحالة منفوخ = *T*، (اللون = الأصفر والحجم = الصغير) أو (العمر = راشد والفعل = الامتداد)، يستلزم جميع متغيرات الخاصية الأربعة اللون، الحجم، العمر، والفعل. فمن الصعب التعبير عن هذا النمط البسيط في شجرة القرار. لا يمكننا استخدام جميع متغيرات الخاصية الأربعة لتقسيم عقدة الجذر. بدلاً من ذلك، علينا اختيار متغير خاصية واحد فقط. ويكون متوسط قيمة مقياس عشوائية المعلومات *(information entropy)* لانفصال ما لتقسيم عقدة الجذر باستخدام كل من متغيرات الخاصية الأربعة هو نفسه تماماً كما هو موضح بالعملية الحسابية أدناه:

$$\text{entropy}(S) = \frac{8}{16}\text{entropy}(D_{Yellow}) + \frac{8}{10}\text{entropy}(D_{Purple})$$
$$= \frac{8}{12} \times \left(-\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8}\right)$$
$$+ \frac{8}{12} \times \left(-\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8}\right)$$
$$= 0.8829$$

نختار عشوائياً اللون = الأصفر *(Color = Yellow)* أو الأرجواني *(Purple)* كمعيار الانفصال لتقسيم عقدة الجذر. يوضح الشكل ٤-٦ شجرة القرار الكاملة لمجموعة بيانات البالون. ويتضح أن شجرة القرار كبيرة بسبعة أنماط للتصنيف مما يؤدي إلى سبع عقد من عقد الورقة، على التوالي:

الشكل ٤-٦
شجرة القرار لمجموعة البيانات الخاصة بالبالون



- Color = Yellow AND Size = Small, with Inflated = T
- Color = Yellow AND Size = Large AND Age = Adult AND Act = Stretch, with Inflated = T
- Color = Yellow AND Size = Large AND Age = Adult AND Act = Dip, with Inflated = F
- Color = Yellow AND Size = Large AND Age = Child, with Inflated = F
- Color = Purple AND Age = Adult AND Act = Stretch, with Inflated = T
- Color = Purple AND Age = Adult AND Act = Dip, with Inflated = F
- Color = Purple AND Age = Child AND, with Inflated = F

- اللون = أصفر والحجم = صغير، مع خاصية منفوخ = T (أي "صحيح").
- اللون = أصفر والحجم = كبير والعمر = راشد، والفعل = ممتد، مع حالة منفوخ = T (أي "صحيح").
- اللون = أصفر والحجم = كبير والعمر = راشد والفعل = منكمش، مع حالة منفوخ = F (أي "خاطئ").
- اللون = أصفر والحجم = كبير والعمر = طفل، مع حالة منفوخ = F (أي "خاطئ ").
- اللون = أرجواني والعمر = راشد والفعل = ممتد، مع حالة منفوخ = T (أي "صحيح").
- اللون = أرجواني والعمر = راشد والفعل = منكمش، مع حالة منفوخ = F (أي "خاطئ").
- اللون = أرجواني والعمر = طفل، مع حالة منفوخ = F (أي "خاطئ").

من ضمن أنماط التصنيف السبعة المذكورة أعلاه، من الصعب أن نرى نمط التصنيف البسيط:

IF (Color = Yellow AND Size = Small) OR (Age = Adult AND Act = Stretch), THEN Inflated = T; OTHERWISE, Inflated = f.

إذا كان (اللون = أصفر، و الحجم = صغير) أو (العمر = راشد و الفعل = ممتد)، إذن تكون خاصية منفوخ = T (أي "صحيح")؛ وإلا تكون خاصية منفوخ = F (أي "خاطئ").

وعلاوةً على ذلك، فإنَّ اختيار معيار الانفصال الأفضل مع متغير خاصية واحد فقط دون النظر إلى تركيب معيار الانفصال هذا مع المعايير اللاحقة وصولاً إلى عقدة الورقة يشبه اتخاذ القرار الأمثل على الصعيد المحلي فقط دون النظر للصعيد الأشمل والأعم. ليس هناك ما يضمن أن اتخاذ القرار الأمثل محلياً في أوقات منفصلة قد يؤدي إلى شجرة القرار الأصغر، أو إلى القرار الأمثل على الصعيد الشامل. بالرغم من ذلك، فإن النظر إلى جميع متغيرات الخاصية وتركيباتها لمعايير وشروط كل انفصال تفضي إلى عملية بحث شاملة لجميع القيم

الممكنة لكل متغيرات الخاصية. وهذا مُكلف حاسوبياً، أو أنه أمر مستحيل أحياناً لمجموعة بيانات كبيرة مع عدد كبير من متغيرات الخاصية.

٤-٦ البرمجيات والتطبيقات (Software and Applications):

يوجد في الموقع الإلكتروني *http://www.knuggets.com* معلومات عن أدوات استكشاف البيانات المختلفة. وحزم البرمجيات التالية تدعم تعلم أشجار القرار والانحدار:

- *Weka (http://www.cs.waikato.ac.nz/ml/weka/)*
- *SPSS AnswerTree (http://www.spss.com/answertree/)*
- *SAS Enterprise Miner (http://sas.com/products/miner/)*
- *IBM Inteligent Miner (http://www.ibm.com/software/data/iminer/)*
- *CART (http://www.salford-systems.com/)*
- *C4.5 (http://www.cse.unsw.edu.au/quinlan)*

بعض التطبيقات الخاصة بأشجار القرار يمكن العثور عليها في (*Ye, 2003*، الفصل ١) و(*Li and Ye, 2001; Ye et al., 2001*).

التمارين **(Exercises) :**

٤-١ قم ببناء شجرة قرار ثنائية لمجموعة بيانات البالون في الجدول ١-١ باستخدام مقياس عشوائية المعلومات *(information entropy)* كمقياس لتجانس البيانات.

٤-٢ قم ببناء شجرة قرار ثنائية لمجموعة بيانات العدسات في الجدول ١-٣ باستخدام مقياس عشوائية المعلومات كمقياس لتجانس البيانات.

٤-٣ قم ببناء شجرة انحدار غير ثنائية لمجموعة البيانات الخاصة بمكوك الفضاء في الجدول ١-٢ باستخدام متغيري الخاصية: درجة حرارة الإطلاق *(Launch Temperature)*، وضغط فحص التسرب *(Leak – Check Pressure)*، ويتم الأخذ بالاعتبار وجود قيمتين نوعيتين لمتغير الخاصية: درجة حرارة الإطلاق، والقيمتان هما:

("منخفضة- *low*" إذا كانت درجة الحرارة < 60، و"طبيعية -*normal*" لدرجات الحرارة الأخرى)؛ أما متغير الخاصية، ضغط فحص التسرب فيكون له ثلاث قيم نوعية هي (50، 100، و200).

٤-٤ قم ببناء شجرة قرار ثنائية أو شجرة قرار غير ثنائية لمجموعة البيانات الموجودة في التمرين ١-١.

٤-٥ قم ببناء شجرة قرار ثنائية أو شجرة قرار غير ثنائية لمجموعة البيانات الموجودة في التمرين ١-٢,

٤-٦ قم ببناء مجموعة بيانات بحيث يكون اختيار الانفصال الأفضل لعقدة الجذر لا يؤدي إلى شجرة القرار الأصغر.

٥- الشبكات العصبية الصناعية للتصنيف والتنبؤ
Artificial Neural Networks For Classification And Prediction

يتم تصميم الشبكات العصبية الصناعية (- *Artificial Neural Networks* *ANNs*) لتحاكي بنية الدماغ البشري من أجل إبداع ذكاء اصطناعي مماثل للذكاء البشري. ومن ثم، فإن الشبكات العصبية الصناعية تستخدم بنية مشابهة للبنية الأساسية للدماغ البشري الذي يتكون من خلايا عصبية وروابط بين الخلايا العصبية. حيث تحتوي الشبكات العصبية الصناعية على وحدات معالجة مشابهة للخلايا العصبية، وروابط بين الوحدات المعالجة. يقدم هذا الفصل نوعين من الشبكات العصبية الصناعية المستخدمة للتصنيف والتنبؤ: الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة *(Perceptron)* والشبكات العصبية الصناعية ذات التغذية الأمامية متعددة الطبقات (*multilayer feedforward ANNs*). في هذا الفصل، نقوم أولاً بوصف وحدات المعالجة، وكيف يمكن استخدام هذه الوحدات لبناء أنواع مختلفة من معماريات الشبكات العصبية الصناعية. نستعرض بعد ذلك الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة، وهي شبكات عصبية صناعية ذات تغذية أمامية أحادية الطبقة، وطريقة تعلم أنماط التصنيف والتنبؤ من خلال الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة. أخيراً، نقوم بوصف الشبكات العصبية الصناعية ذات التغذية الأمامية متعددة الطبقات، ثم وصف خوارزمية التعلم بالتوالد الخلفي (*back-propagation learning algoritnn*). سيتم استعراض حزم من قائمة البرمجيات التي تدعم الشبكات العصبية الصناعية. كما سيتم استعراض بعض تطبيقات الشبكات العصبية الصناعية مع المراجع الخاصة بها.

٥-١ وحدات المعالجة للشبكات العصبية الصناعية (Processing Units of ANNs):

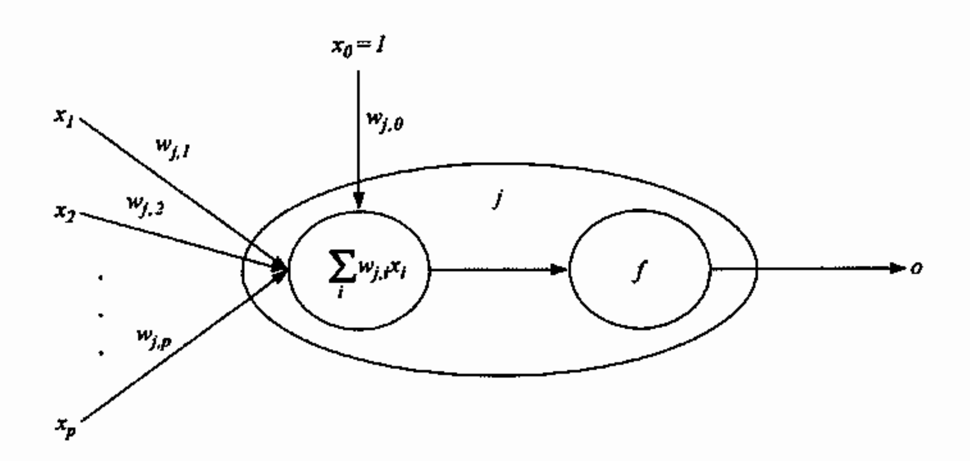
يوضح الشكل ٥-١ إحدى وحدات المعالجة في شبكة عصبية صناعية (*ANN*)، وهي الوحدة *j*، حيث تأخذ هذه الوحدة عدد p من المدخلات، $x_1 x_2$ ،...، x_p ومُدخَلة خاصة أخرى، $x_0 = 1$ وتنتج مُخرَجة واحدة هي، o. حيث يتم استخدام المدخلات، x_p ،...، x_2 وx_1 ، والمخرجة o، لتمثيل المدخلات والمخرجات الخاصة بمسألة أو مشكلة معينة. لنأخذ

مثالاً من مجموعة البيانات الخاصة بمكوك الفضاء في الجدول ١-٢. قد يكون لدينا المتغيرات x_1 وx_2 وx_3 لتمثيل درجة حرارة الإطلاق (*Launch Temperature*)، وضغط فحص التسرب (*Leak–Check Pressure*) ، والترتيب الزمني للرحلة (*Temporal Order of Flight*)، على التوالي، ويكون المتغير o لتمثيل عدد الحلقات الدائرية ذات الأحمال الثقيلة (*O-Rings with Stress*). والمدخلة x_0 عبارة عن جزء لا يتجزأ لكل وحدة من وحدات المعالجة، وهي تأخذ القيمة واحد دائماً. كل مُدخَل من المدخلات، x_i يرتبط بالوحدة j مع وزن الرابط $w_{j,i}$. ويسمى وزن الرابط $w_{j,0}$ بالتحيز (*bias*)، أو الحد (*threshold*)، وذلك لسبب سيتم توضيحه لاحقاً. تقوم الوحدة j بمعالجة المدخلات عن طريق إيجاد صافي المجموع أولاً، وهو المجموع الموزون للمدخَلات، وذلك على النحو التالي:

$$net_j = \sum_{i=0}^{p} w_{j,i}\, x_i \qquad (1\text{-}5)$$

الشكل (٥-١)

وحدة معالجة بالشبكة العصبية الصناعية (*ANN*)



لتكن المتجهات x وw معرّفة على النحو التالي:

$$x = \begin{bmatrix} x_0 \\ \vdots \\ x_p \end{bmatrix} \quad w' = \begin{bmatrix} w_{j,0} & \cdots & w_{j,p} \end{bmatrix}$$

يمكن تمثيل المعادلة ٥-١ على النحو التالي:

$$net_j = w'x. \qquad (٥-٢)$$

ثم تقوم الوحدة، j، بتطبيق دالة تحول، f، إلى صافي المجموع وتوجد الناتج أو المخرَجة، o، على النحو التالي:

$$o = f(net_j). \qquad (٥-٣)$$

فيما يلي يتم استعراض خمس دوال من دوال التحول الشائعة، ويتم توضيحها في الشكل ٥-٢:

١- دالة الإشارة (*Sign function*):

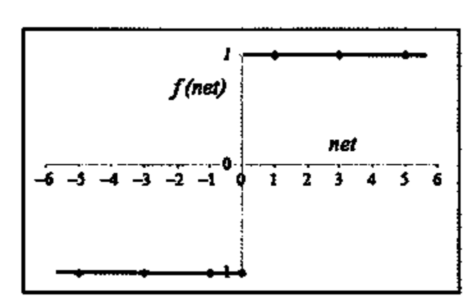
$$o = sgn(net) = \begin{cases} 1 & \text{if } net > 0 \\ -1 & \text{if } net \le 0 \end{cases} \qquad (٥-٤)$$

٢- دالة الحد الثابت (*Hard limit Function*):

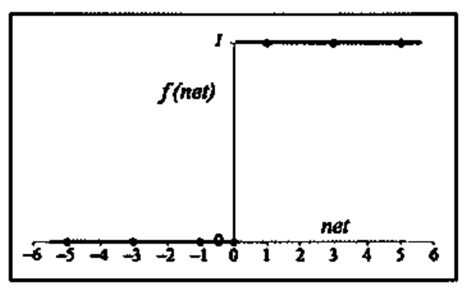
$$o = hardlim(net) = \begin{cases} 1 & \text{if } net > 0 \\ 0 & \text{if } net \le 0 \end{cases} \qquad (٥-٥)$$

الشكل (٥-٢)

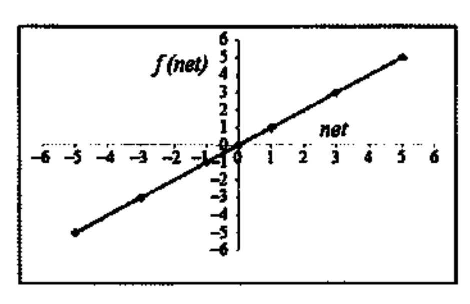
أمثلة على دوال التحول



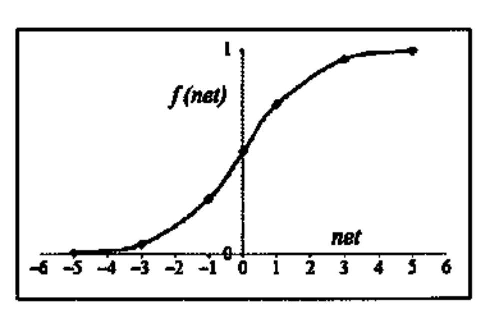
The Sign function – دالة الإشارة



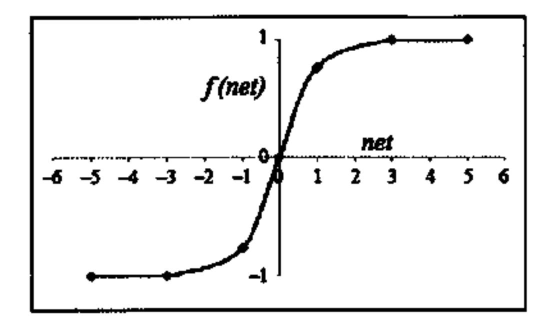
The hard limit function - دالة الحد الثابت



The linear function - الدالة الخطية



الدالة السينية (على شكل حرف *S*)
The sigmoid function



The hyperbolic tangent function - دالة الظل القطعي

٣- الدالة الخطية: (*Linear Function*):

$$o = lin(net) = net \qquad (٥-٦)$$

٤- الدالة السينية: (*Sigmoid function*):

$$o = sig(net) = \frac{1}{1 + e^{-net}} \qquad (٥-٧)$$

٥- دالة الظل القطعي: (*Hyperbolic tangent function*):

$$o = tanh(net) = \frac{e^{net} - e^{-net}}{e^{net} + e^{-net}}. \qquad (٥-٨)$$

من خلال المعطيات التالية الخاصة بمتجه المدخَلات ومتجهة وزن الارتباط (w')

$$x = \begin{bmatrix} 1 \\ 5 \\ -6 \end{bmatrix} \quad w' = [-1.2 \quad 3 \quad 2],$$

يتم احتساب ناتج الوحدة لكل من دوال التحول الخمسة المذكورة آنفاً على النحو التالي:

$$net = w'x = [-1.2 \quad 3 \quad 2]\begin{bmatrix} 1 \\ 5 \\ -6 \end{bmatrix} = 1.8$$

$$o = sgn(net) = 1$$

$$o = hardlim(net) = 1$$

$$o = lin(net) = 1.8$$

$$o = sig(net) = 0.8581$$

$$o = tanh(net) = 0.9468.$$

تكفي وحدة معالجة واحدة لتنفيذ الدالة AND المنطقية. حيث يعطي الجدول ٥-١ المدخلات والمخرجات للدالة AND، وأربعة سجلات للبيانات الخاصة بهذه الدالة. الدالة AND تحتوي على قيم المخرجات 1- و 1. الشكل ٥-٣ يوضح تطبيق الدالة AND باستخدام وحدة معالجة واحدة.

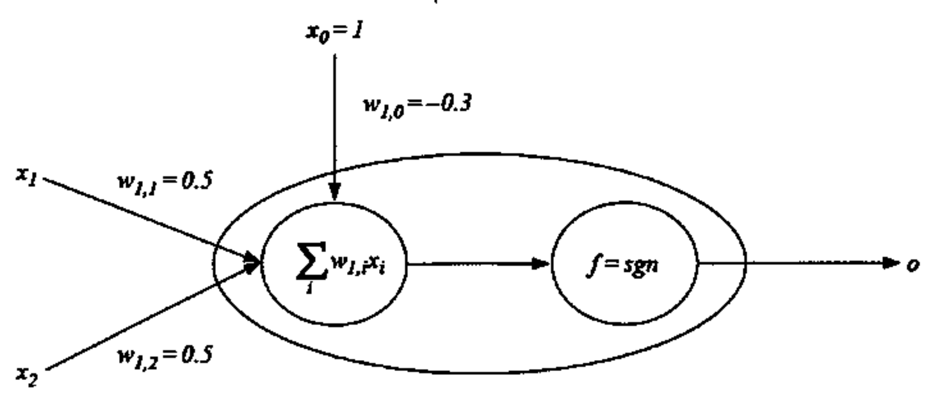
الجدول (٥-١)

الدالة AND

x_1	x_2	o
Inputs - المدخلات		**Output - المخرجات**
-1	-1	-1
-1	1	-1
1	-1	-1
1	1	1

الشكل (٥-٣)

تطبيق الدالة AND باستخدام وحدة معالجة واحدة



من بين دوال التحول الخمس في الشكل ٥-٢، يمكن لدالة الإشارة ودالة الظل القطعي أن ينتج عنهما مجموعة من قيم المخرجات التي تتراوح بين 1- إلى 1. يتم استخدام دالة الإشارة كدالة تحول لوحدة المعالجة لتطبيق دالة AND تتطلب أول ثلاثة سجلات بيانات قيمة المخرجات 1-. ينبــغي أن يكــون المجمــوع الموزون لمدخلات سجلات البيانات الثلاثة الأولى، $w_{1,0}\,x_0 + w_{1,1}x_1 + w_{1,2}x_2$، في النطاق [-$1$، ٠]. ويتطلب سجل البيانات الأخير قيمة المخرجات التي تبلغ 1، وينبغي أن يكون المجموع الموزون للمدخلات في النطاق $(0,1]$. ويجب أن يكون وزن الارتباط $w_{1,0}$ ، ذا قيمة سالبة لجعل net لأول ثلاثة سجلات من سجلات البيانات أقل من الصفر، وأيضاً لجعل net لآخر سجلات بيانات أكبر من الصفر. ومن ثم، فإن وزن الارتباط $w_{1,0}$ ، يكون بمثابة الحد (الحاجز) أمام المجموع الموزون للمدخلات لجعل قيمة net أكبر من أو أقل من الصفر. وهذا هو السبب في أن وزن الارتباط لـ $x_0=1$ يُدعى بالحد (الحاجز) أو التحيز. في الشكل ٥-٣، تمّ وضع قيمة $w_{1,0}$ عند -0.3، يمكن أن يتم تمثيل المعادلة ٥-١ على النحو التالي لإظهار دور الحد (الحاجز) أو التحيز، b:

$$net = w'x + b, \qquad (٥\text{-}٩)$$

حيث:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \quad w' = \begin{bmatrix} w_{j,1} & \cdots & w_{j,p} \end{bmatrix}.$$

ويتضح تالياً حساب قيمة المخرجات لكل مدخل من المدخلات الموضحة في الجدول ٥ – ١:

$$o = sgn(net) = sgn\left(\sum_{i=0}^{2} w_{1,i}x_i\right) = sgn[-0.3 \times 1 + 0.5 \times (-1) + 0.5 \times (-1)]$$
$$= sgn(-0.3 - 1) = sgn(-1.3) = -1$$

$$o = sgn(net) = sgn\left(\sum_{i=0}^{2} w_{1,i}x_i\right) = sgn[-0.3 \times 1 + 0.5 \times (-1) + 0.5 \times (1)]$$
$$= sgn(-0.3 + 0) = sgn(-0.3) = -1$$

$$o = sgn(net) = sgn\left(\sum_{i=0}^{2} w_{1,i}x_i\right) = sgn[-0.3 \times 1 + 0.5 \times (1) + 0.5 \times (-1)]$$
$$= sgn(-0.3 + 0) = sgn(-0.3) = -1$$

$$o = sgn(net) = sgn\left(\sum_{i=0}^{2} w_{1,i}x_i\right) = sgn[-0.3 \times 1 + 0.5 \times (1) + 0.5 \times (1)]$$
$$= sgn(-0.3 + 1) = sgn(0.7) = 1$$

يعطي الجدول ٥-٢ المدخلات والمخرجات الخاصة بالدالة OR المنطقية. ويبين الشكل ٥-٤ تطبيق الدالة OR باستخدام وحدة معالجة واحدة.

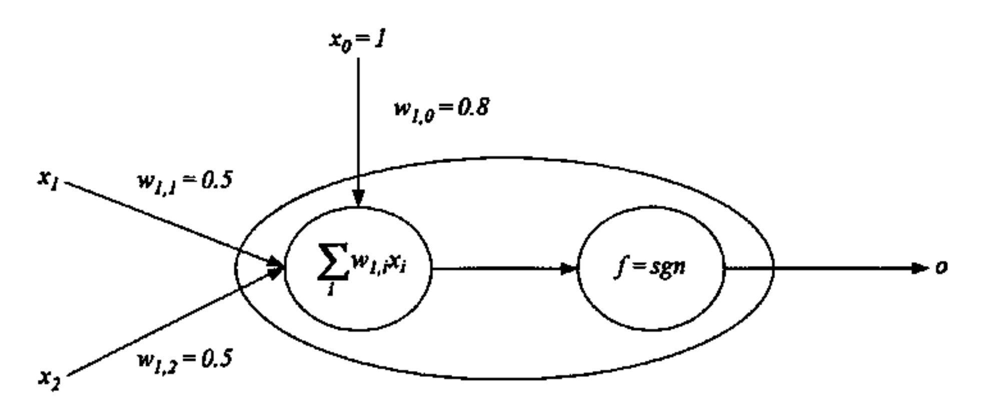
الجدول (٥-٢)

الدالة OR

Output - المخرجات	Inputs - المدخلات	
o	x_2	x_1
-1	-1	-1
1	1	-1
1	-1	1
1	1	1

الشكل (٥-٤)

تطبيق الدالة OR باستخدام وحدة معالجة واحدة



يتطلب سجل البيانات الأول فقط قيمة المخرجات 1- ، وتتطلب سجلات البيانات الثلاثة الأخرى أن تكون قيمة المخرجات 1. يعطي سجل البيانات الأول فقط المجموع الموزون 1- من المدخلات، وتعطي سجلات البيانات الثلاثة الأخرى المجموع الموزون للمدخلات في النطاق [$-0.5, 1$]. ومن ثم، فإن أي قيمة للحد (الحاجز) $w_{1,0}$ في النطاق ($0.5, 1$) ستجعل قيمة *net* لسجل البيانات الأول أقل من الصفر، وجعل قيمة *net* لسجلات البيانات الثلاثة الأخيرة أكبر من الصفر.

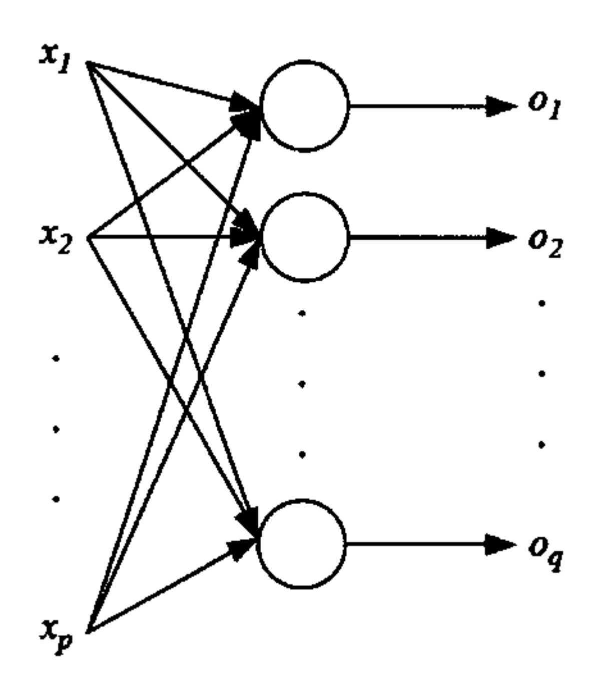
٥-٢ معماريات الشبكات العصبية الصناعية (Architectures of ANNs):

يمكن استخدام وحدات معالجة الشبكات العصبية الصناعية (*ANNs*) لبناء أنواع مختلفة من معماريات الشبكات العصبية الصناعية (*ANNs*). نستعرض تصميمين أو معماريتين للشبكات العصبية الصناعية (*ANNs*): الشبكات العصبية الصناعية ذات التغذية الأمامية (*Feed forward ANNS*)، والشبكات العصبية الصناعية الدورية (*Recurrent ANNs*). يتم استخدام الشبكات العصبية الصناعية ذات التغذية الأمامية على نطاق واسع. ويبين الشكل ٥-٥ الشبكات العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة وكاملة الترابط، والتي يرتبط فيها مدخل من المدخلات بكل وحدة من

وحدات المعالجة. ويبين الشكل ٥-٦ الشبكات العصبية الصناعية ذات التغذية الأمامية ثنائية الطبقات والكاملة الترابط.

الشكل (٥-٥)

معمارية الشبكات العصبية الصناعية ذات التغذية الأمامية الأحادية الطبقة

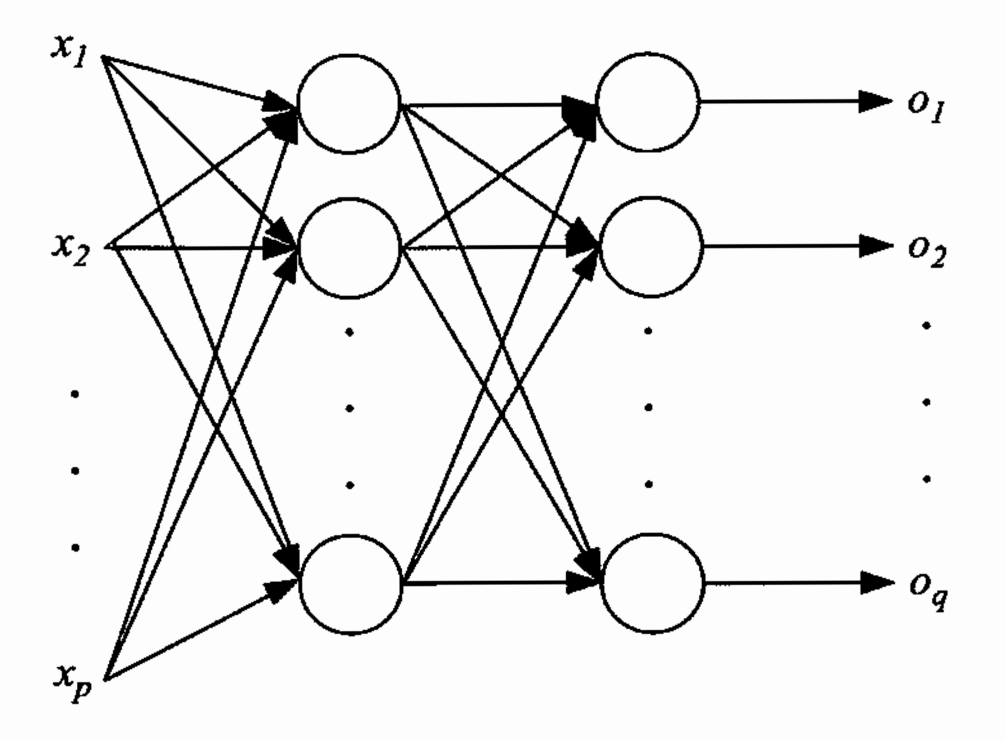


يُلاحظ أن المدخلة x_0 لكل وحدة من وحدات المعالجة لا تظهر بشكل صريح في معماريات الشبكات العصبية الصناعية *ANN* في الأشكال ٥-٥ و٥-٦. تحتوي الشبكات العصبية الصناعية ذات التغذية الأمامية *ANN* ثنائية الطبقات في الشكل ٥-٦ على طبقة مخرجات لوحدات المعالجة لإنتاج المخرجات، وطبقة مخفية لوحدات المعالجة التي تشكل مخرجاتها مدخلات لوحدات المعالجة في طبقة المخرجات. يتم ربط كل مدخل من المدخلات بكل وحدة من وحدات المعالجة في الطبقة المخفية، ويتم ربط كل وحدة من وحدات المعالجة في الطبقة المخفية بكل وحدة من وحدات المعالجة في طبقة المخرجات. في الشبكات العصبية الصناعية ذات التغذية الأمامية *ANN* لا يوجد روابط عكسية بين

وحدات المعالجة، بمعنى آخر، لا يتم استخدام مخرجات وحدة معالجة معينة ليكون جزءاً من المدخلات لنفس وحدة المعالجة بشكل مباشر أو غير مباشر. ليس بالضرورة أن تكون الشبكات العصبية الصناعية *ANN* مترابطة ترابطاً كاملاً كما هو الحال في الأشكال ٥-٥، و٥-٦. قد تستخدم وحدات المعالجة نفس دالة التحول، أو دوال تحول مختلفة.

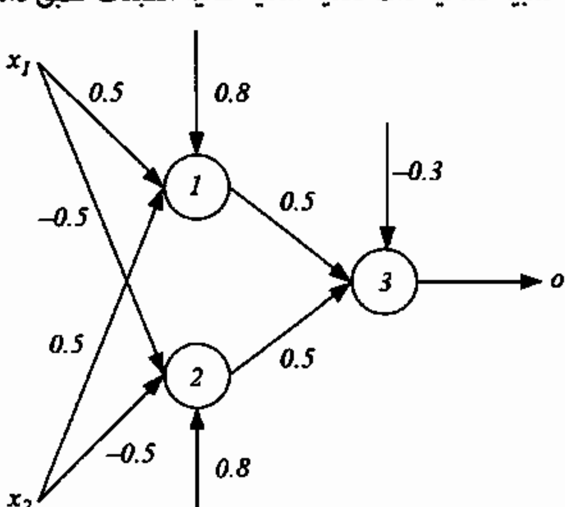
الشكل (٥-٦)

معمارية الشبكات العصبية الصناعية ذات التغذية الأمامية الثنائية الطبقات



الشكل (٥-٧)

شبكات عصبية صناعية ذات تغذية أمامية ثنائية الطبقات تطبق دالة *XOR*



الشبكات العصبية الصناعية *ANNs* في الأشكال ٥-٣ و٥-٤، على التوالي، هي أمثلة على الشبكات العصبية الصناعية ذات التغذية الأمامية الأحادية الطبقة. ويبين الشكل ٥-٧ الشبكات العصبية الصناعية ذات التغذية الأمامية ثنائية الطبقة كاملة الترابط مكونة من طبقة مخفية واحدة تحتوي وحدتي معالجة، وطبقة مخرجات تحتوي وحدة معالجة واحدة لتنفيذ الدالة المنطقية والحصرية *OR*، ويُرمَز لها بالرمز (*XOR*). يوضح الجدول ٥-٣ المدخلات والمخرجات الخاصة بالدالة *XOR*.

إن عدد المدخلات، وعدد المخرجات في الشبكات العصبية الصناعية *ANN* يعتمد على الدالة المستخدمة من قبل الشبكات العصبية الصناعية *ANN* على سبيل المثال، فإن الدالة *XOR* لها مدخلان اثنان ومُخرَج واحد ومن ثم يمكن تمثيلها بشبكة عصبية صناعية *ANN* تحتوي مُدخَلين اثنين ومُخرِج واحد، على التوالي. غالباً ما يتم تحديد عدد وحدات المعالجة في الطبقة المخفية، والتي تُسمى بالوحدات المخفية، تجريبياً بحيث تأخذ في الاعتبار درجة تعقيد الدالة التي تقوم الشبكات العصبية الصناعية *ANN* باستخدامها. بشكل عام، كلما كانت الدالة أكثر تعقيداً، كانت هناك حاجة إلى المزيد من الوحدات المخفية. شبكات الـ *ANN* ذات

التغذية الأمامية ثنائية الطبقات مع دالة سينية أو دالة الظل القطعي يكون لها من القدرة على تطبيق حالة معطاة (*Witten et al., 2011*).

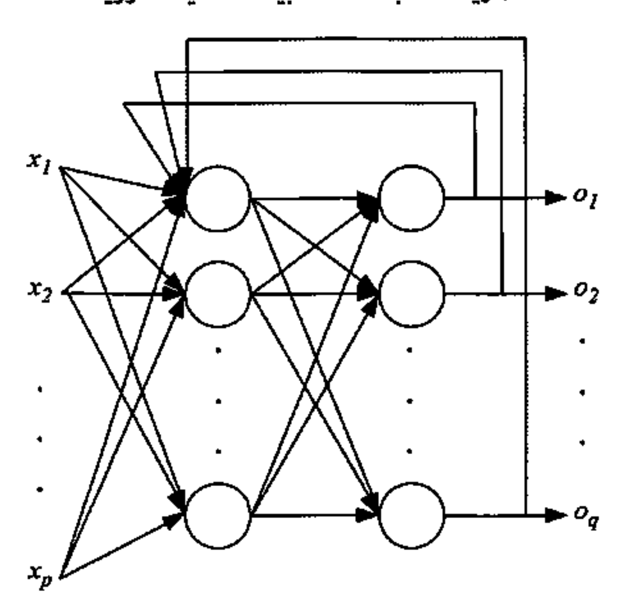
الجدول (٥-٣)

الدالة *XOR*

المخرجات - Output	المدخلات - Inputs	
o	x_2	x_1
-1	-1	-1
1	1	-1
1	-1	1
-1	1	1

الشكل (٥-٨)

معماريات الشبكات العصبية الصناعية الدورية



ويبين الشكل ٥-٨ معمارية الشبكات العصبية الصناعية الدورية مع روابط عكسية تستخدم المخرجات على هيئة مدخلات إلى الوحدة المخفية الأولى (ظاهرة) ووحدات مخفية أخرى (غير ظاهرة). تسمح الروابط العكسية للشبكات العصبية الصناعية *ANN* بالتقاط السلوك الزمني، بحيث أن المخرجات في الوقت $t + 1$ تعتمد على المخرجات، أو على حالة شبكات الـ *ANN* في الوقت t. ومن ثم، فإن شبكات الـ *ANN* الدورية مثل تلك الموضحة في الشكل ٥-٨ تحتوي روابط عكسية لالتقاط السلوكيات الزمنية.

٥-٣ طرق تحديد أوزان الروابط في الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة

(Methods of Determining Connection Weights for a Perceptron):

لاستخدام شبكة الـ *ANN* لتطبيق دالة ما، علينا أولاً تحديد معمارية شبكة الـ *ANN*، بما في ذلك عدد المدخلات، وعدد المخرجات، وعدد الطبقات، وعدد وحدات المعالجة في كل طبقة، ودالة التحول لكل وحدة من وحدات المعالجة. ثم تحتاج لتحديد أوزان الروابط. في هذا الجزء، نقوم بوصف طريقة بيانية، وطريقة تعلُمية لتحديد أوزان الروابط لشبكة الـ *Perceptron*، وهي شبكة عصبية صناعية ذات تغذية أمامية أحادية الطبقة مع دالة الإشارة (*sign function*)، أو دالة تحول الحد الثابت (*hard limit transfer function*). على الرغم من أنه يتم شرح المفاهيم والأساليب في هذا الجزء باستخدام دالة تحول الإشارة لكل وحدة من وحدات المعالجة في شبكة الـ *perception* ، فإن هذه المفاهيم والأساليب قابلة للتطبيق أيضاً على شبكة الـ *perceptron* مع دالة تحول الحد الثابت لكل وحدة من وحدات المعالجة.

في الجزء ٥-٤، نستعرض طريقة التعلم بالتوالد الخلفي لتحديد أوزان الروابط للشبكات العصبية الصناعية ذات التغذية الأمامية المتعددة الطبقات.

٥-٣-١ الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة (Perceptron)

يتم استخدام الرموز التالية لتمثيل الشبكة العصبية الصناعية ذات التغذية الأمامية الأحادية الطبقة والمرتبطة ارتباطاً كاملاً بعدد مدخلات p، ووحدات معالجة في طبقة المخرجات بغرض إنتاج مخرجات عددها q، ودالة تحول الإشارة لكل وحدة من وحدات المعالجة، كما هو مبين في الشكل ٥-٥:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \quad o = \begin{bmatrix} o_1 \\ \vdots \\ o_q \end{bmatrix} \quad w' = \begin{bmatrix} w_{1,1} & \cdots & w_{1,p} \\ \vdots & \ddots & \vdots \\ w_{q,1} & \cdots & w_{q,p} \end{bmatrix} = \begin{bmatrix} w'_1 \\ \vdots \\ w'_q \end{bmatrix} \quad w_j = \begin{bmatrix} w_{j,1} \\ \vdots \\ w_{j,p} \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_q \end{bmatrix}$$

$$o = sgn(w'x + b). \qquad (١٠-٥)$$

٥-٣-٢ خصائص وحدة المعالجة (Properties of a Processing Unit):

بالنسبة لوحدة معالجة معينة j، فإنّ المخرجات $o=sgn(net)=sgn(w'_j\ x+b_j)$ تفصل متجهات المدخلات، xs إلى منطقتين: منطقة يكون بها $net>0$، و $o=1$، والمنطقة الأخرى يكون بها $net\leq 0$، و $o=-1$.

إن المعادلة، $net = w'_j\ x + b_j = 0$، هي حد القرار (*decision boundary*) في فضاء المدخلات التي تفصل بين المنطقتين. على سبيل المثال، قيم x معطاة في فضاء ثنائي الأبعاد، والوزن، والتحيز التالية:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad w'_j = \begin{bmatrix} -1 & 1 \end{bmatrix} \quad b_j = -1,$$

حد القرار هو:

$$w'_j x + b_j = 0$$
$$-x_1 + x_2 - 1 = 0$$
$$x_2 = x_1 + 1.$$

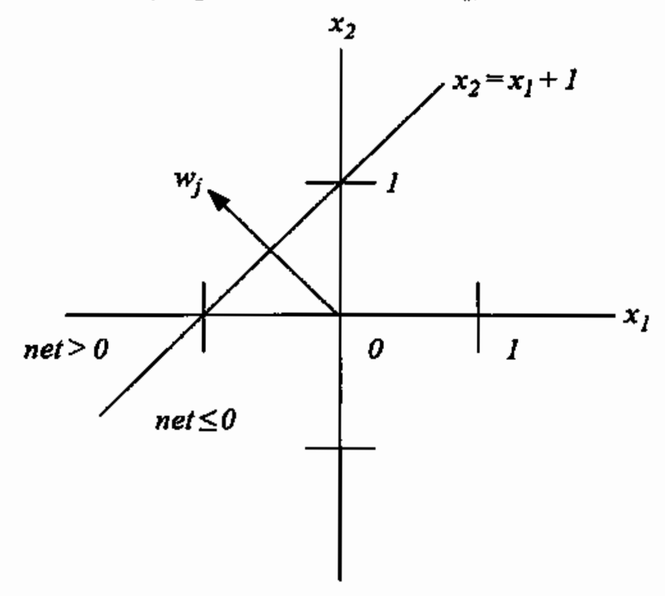
ويوضح الشكل ٥-٩ حد القرار، وفصل فضاء المدخلات إلى منطقتين بواسطة حد القرار. الميل (*slope*) ونقطة التقاطع (*intercept*) للخط الذي يمثل حد القرار في الشكل ٥-٩، هما:

$$\text{slope} = \frac{-w_{j,1}}{w_{j,2}} = \frac{1}{1} = 1$$

$$\text{intercept} = \frac{-b_j}{w_{j,2}} = \frac{1}{1} = 1.$$

الشكل (٥-٩)

مثال على حد القرار وفصل بين فضاء المدخلات إلى منطقتين من خلال وحدة المعالجة



كما هو موضح في الشكل ٥-٩، تتميز وحدة المعالجة بالخصائص التالية:

- يكون متجه الوزن متعامداً على حد (حاجز) القرار.
- يشير متجه الوزن إلى الجانب الموجب $(net > 0)$ لحد القرار.
- الموقع الخاص بحد القرار يمكن إزاحته من خلال تغيير b. إذا كانت $b=0$، فإن حد القرار يمر من خلال نقطة الأصل، على سبيل المثال نقطة الأصل هي $(0، 0)$ في الفضاء ثنائي الأبعاد.

- لأن حد القرار عبارة عن معادلة خطية، يمكن لوحدة المعالجة أن تقوم بتطبيق دالة قابلة للفصل خطياً فقط.

تُستخدم هذه الخصائص لوحدة المعالجة في الطريقة البيانية لتحديد أوزان الروابط في الجزء ٥-٣-٣، وطريقة التعلم لتحديد أوزان الروابط في الجزء ٥-٣-٤.

٥-٣-٣ الأسلوب البياني لتحديد أوزان الروابط والتحيزات

(Graphical Method of Determining Connection Weights and Biases):

يتم الأخذ بالخطوات التالية كأسلوب بياني لتحديد أوزان الروابط للشبكة العصبية الصناعية ذات التغذية الأمامية الأحادية الطبقة *(perception)* بعدد مدخلات p ومخرج واحد، ووحدة معالجة واحدة لإنتاج المخرجات، ودالة تحول الإشارة لوحدة المعالجة:

١- ارسم نقاط البيانات لسجلات البيانات في مجموعة البيانات التدريبية (الاستكشافية) لهذه لدالة.

٢- ارسم حد القرار لفصل نقاط البيانات ذات القيم $o=1$ عن نقاط البيانات ذات القيم $o=-1$.

٣- ارسم متجه الوزن واجعله متعامداً على حد القرار، ويشير إلى الجانب الموجب من حد القرار. وتحدد إحداثيات متجه الوزن أوزان الروابط.

٤- استخدم إحدى الطريقتين التاليتين لتحديد التحيز b:

أ- استخدم نقطة تقاطع مستقيم حد القرار مع أوزان الروابط لتحديد التحيز(b).

ب- اختر عدداً قليلاً من نقاط البيانات على كلا الجانبين الموجب والسالب لمستقيم حد القرار بحيث تكون النقاط هي الأقرب إلى مستقيم حد القرار واستخدم نقاط البيانات تلك وأوزان الروابط لتحديد التحيز (b).

هذه الخطوات موضحة في المثال ٥-١.

المثال (٥-١)

استخدم الطريقة البيانية لتحديد أوزان الروابط للشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة *(perceptron)* المحتوية على وحدة معالجة واحدة للدالة *AND* في الجدول ٥-١.

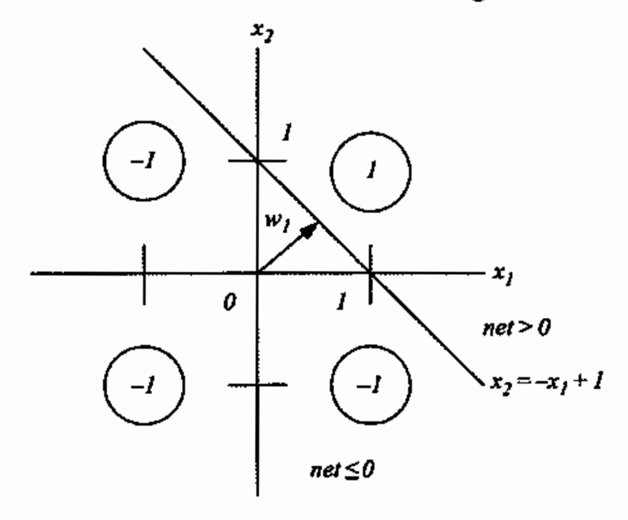
في الخطوة 1، قمنا برسم الدوائر الأربعة في الشكل ٥-١٠ لتمثل نقاط البيانات الأربعة للدالة *AND* وقد تمّ إبراز قيمة المخرجات لكل نقطة من نقاط البيانات داخل دائرة لنقطة. في الخطوة ٢، نستخدم معادلة حد القرار،

$x_2 = -x_1 + 1$، لفصل نقاط البيانات الثلاثة التي بها $o=-1$ عن نقطة البيانات التي بها $o = 1$. نقطة تقاطع مستقيم حد القرار هي 1 بحيث تكون $x_2 = 1$ عند وضع x_1 عند صفر. في الخطوة 3، رسمنا متجه الوزن $w_1 = (0.5, 0.5)$، وهو متعامد على مستقيم حد القرار ويشير إلى الجانب الموجب منه. ومن ثم، يكون لدينا $w_{1,1} = 0.5$، $w_{1,2} = 0.5$. في الخطوة 4، نقوم باستخدام المعادلة التالية لتحديد التحيز:

$$w_{1,1}x_1 + w_{1,2}x_2 + b = 0$$
$$w_{1,2}x_2 = -w_{1,1}x_1 - b$$

الشكل (٥-١٠)

توضيح الطريقة البيانية لتحديد أوزان الروابط



$$\text{intercept} = -\frac{b}{w_{1,2}}$$

$$1 = -\frac{b}{0.5}$$

$$b = -0.5.$$

وإذا ما حركنا مستقيم حد القرار بحيث تكون نقطة التقاطع عند 0.6، فإننا نحصل على $b=-0.3$، وبالضبط على نفس الشبكة العصبية الصناعية ANN للدالة AND كما هو مبين في الشكل ٥-٣. باستخدام طريقة أخرى في الخطوة ٤، نختار نقطة البيانات $(1، 1)$ على الجانب الموجب لمستقيم حد القرار، ونقطة البيانات $(1، -1)$ على الجانب السالب لحد القرار، وأوزان الروابط $w_{1,1}=0.5,\ w_{1,2}=0.5$، لتحديد التحيز b على النحو التالي:

$$net = w_{1,1}x_1 + w_{1,2}x_2 + b$$

$$net = 0.5 \times 1 + 0.5 \times 1 + b > 0$$

$$b > -1$$

و

$$net = w_{1,1}x_1 + w_{1,2}x_2 + b$$

$$net = 0.5 \times (-1) + 0.5 \times 1 + b \leq 0$$

$$b \leq 0.$$

ومن ثم يكون لدينا:

$$-1 < b \leq 0.$$

بجعل $b=-0.3$، نحصل على نفس الشبكة العصبية الصناعية ANN للدالة AND كما هو مبين في الشكل ٥-٣. إن شبكة الـ ANN بالأوزان، والتحيز، وحد القرار كما هو الحال في

الشكل ٥-١٠ ينتج عنها المخرجات الصحيحة للمدخلات في كل سجل من سجلات البيانات الواردة في الجدول ٥-١. للشبكة *ANN* أيضاً القدرة على تعميم تصنيف أي متجه من متجهات المدخلات على الجانب السالب لحد القرار إلى $o=-1$، وأي متجه من متجهات المدخلات على الجانب الموجب من حد القرار إلى $o=1$. بالنسبة للشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة المحتوية على وحدات مخرجات متعددة، يتم تطبيق الطريقة البيانية لتحديد أوزان الروابط والتحيز لكل وحدة من وحدات المدخلات.

٥-٣-٤ طريقة تعلُم لتحديد أوزان الروابط والتحيزات
(Learning Method of Determining Connection Weights and Biases):

نستخدم السجلين التاليين من سجلات البيانات الأربعة للدالة *AND* في مجموعة البيانات التدريبية لتوضيح طريقة تعلم تحديد أوزان الروابط للشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة المحتوية على وحدة معالجة واحدة بدون تحيز:

1. $x_1=-1$ $x_2=-1$ $t_1=-1$
2. $x_1=1$ $x_2=1$ $t_1=1,$

حيث تشير t_1 إلى المخرجات المستهدفة لوحدة المعالجة 1 التي تحتاج إلى أن يتم إنتاجها لكل سجل من سجلات البيانات. يتم رسم سجلي بيانات في الشكل ٥-١١.

نقوم بإعطاء قيم أولية لأوزان الروابط باستخدام قيم عشوائية، $w_{1,1}(k)=-1$ و $w_{1,2}(k)=0.8$، حيث تشير k إلى عدد التكرار عندما يتم إسناد الأوزان أو تحديثها. في البداية، تكون $k=0$. نقدم مدخلات أول سجل بيانات إلى الشبكة العصبية الصناعية ذات التغذية الأمامية الأحادية الطبقة بوحدة معالجة واحدة:

$$net = w_{1,1}(0)\,x_1 + w_{1,2}(0)\,x_2 = (-1)\times(-1) + 0.8\times(-1) = -1.8.$$

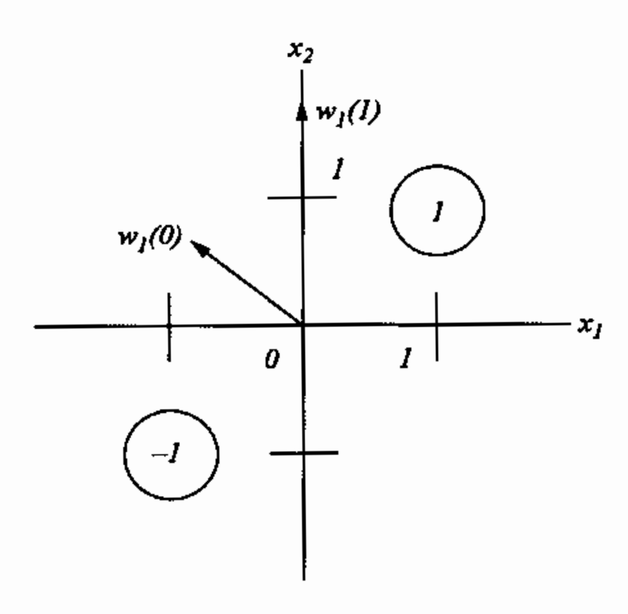
وحيث إن $net<0$ ، فيكون $o_1=-1$. ومن ثم، فإن شبكة الـ *percepton* مع متجه الوزن $(-1,0.8)$ تنتج المخرجات المستهدفة لمدخلات أول سجل بيانات، $t_1=-1$. ليست هناك

حاجة لتغيير أوزان الروابط. بعد ذلك، نقوم بتقديم مدخلات سجل البيانات الثاني إلى شبكة الـ *perceptron*:

$$net = w_{1,1}(0)\, x_1 + w_{1,2}(0)\, x_2 = (-1) \times 1 + 0.8 \times 1 = -0.2.$$

وحيث إن $net<0$، فيكون $o_1= -1$، والذي يختلف عن المخرجات المستهدفة لسجل البيانات هذا،$t_1=1$. ومن ثم، يجب أن يتم تغيير أوزان الروابط من أجل إنتاج المخرجات المستهدفة.

الشكل (٥-١١)
توضيح طريقة تعلم تغيير أوزان الروابط



وتُستخدَم المعادلات التالية لتغيير أوزان الروابط لوحدة المعالجة j:

$$\Delta w_j = \frac{1}{2}\left(t_j - o_j\right)x \qquad (٥\text{-}١١)$$

$$w_j(k+1) = w_j(k) + \Delta w_j. \qquad (١٢-٥)$$

في المعادلة ٥-١١، إذا كانت قيمة $(t - o)$ صفراً، بمعنى، $t = 0$ ، فإنه لا يكون هناك أي تغيير في الأوزان. إذا كانت، $o= -1$،

$$\Delta w_j = \frac{1}{2}\left(t_j - o_j\right)x = \frac{1}{2}(1-(-1))x = x.$$

بإضافة x إلى $w_j(k)$، مما يعني، حل $w_j(k)+x$ في المعادلة ٥-١٢، نحرك متجه الوزن بالقرب من x، ونجعل نقطة متجه الوزن باتجاه x بشكل أكبر لأننا نريد أن يشير متجه الوزن إلى الجانب الموجب من حد القرار، وأن تقع على x على الجانب الموجب من حد القرار. إذا كان $t_1=-1$ ، و $o_1=-1$،

$$\Delta w_j = \frac{1}{2}\left(t_j - o_j\right)x = \frac{1}{2}(-1-1)x = -x.$$

بطرح x من $w_j(k)$، مما يعني، حل $w_j(k) - x$ في المعادلة ٥-١٢، نحرك متجه الوزن بعيداً عن x، ونجعل نقطة متجه الوزن أقرب إلى الاتجاه المعاكس لـ x، لأن x تقع على الجانب السالب من حد القرار مع $t=-1$ ، ونريد متجه الوزن أن يشير في النهاية إلى الجانب الموجب من حد القرار.

باستخدام المعادلات ٥-١١، و٥-١٢، نقوم بتحديث أوزان الروابط استناداً إلى المدخلات والمخرجات المستهدفة والفعلية لسجل البيانات الثاني، وذلك على النحو التالي:

$$\Delta w_1 = \frac{1}{2}(t_1 - o_1)x = \frac{1}{2}\left(1-(-1)\right)\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix}$$

$$w_1(1) = w_1(0) + \Delta w_1 = \begin{bmatrix}-1\\0.8\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}0\\1.8\end{bmatrix}.$$

يظهر متجه الوزن الجديد، $w_1(1)$، في الشكل ٥-١١. كما هو واضح من الشكل ٥-١١، فإن $w_1(1)$، تظهر أقرب إلى سجل البيانات الثاني x من $w_1(0)$، وتشير بشكل أكبر إلى اتجاه x لأن x يكون لديها $t = 1$ ، ومن ثم تقع على الجانب الموجب من حد القرار.

مع الأوزان الجديدة، نقوم باستعراض مدخلات سجلات البيانات إلى شبكة الـ *perceptron* مرة أخرى في التكرار الثاني لتقييم وتحديث الأوزان إذا لزم الأمر. ونستعرض مدخلات أول سجل بيانات:

$$net = w_{1,1}(1)\,x_1 + w_{1,2}(1)\,x_2 = 0 \times (-1) + 1.8 \times (-1) = -1.8.$$

وحيث إن $net < 0$، يكون لدينا $o_1 = -1$. ومن ثم، فإن شبكة الـ *perceptron* بمتجه الوزن $(0, 1.8)$ تنتج المخرجات المستهدفة لمدخلات أول سجل بيانات، $t_1 = -1$. ومع $(t_1 - o_1) = 0$، ليست هناك حاجة لتغيير أوزان الروابط. بعد ذلك نقوم باستعراض مدخلات سجل البيانات الثاني إلى شبكة الـ *perceptron*:

$$net = w_{1,1}(1)\,x_1 + w_{1,2}(1)\,x_2 = 0 \times 1 + 1.8 \times 1 = 1.8.$$

وحيث إن $net > 0$، لدينا $o_1 = 1$. ومن ثم، فإن الشبكة العصبية شبكة الـ *perceptron* بمتجه الوزن $(0, 1.8)$ تنتج المخرجات المستهدفة لمدخلات سجل البيانات الثاني، $t = 1$. مع $(t - o) = 0$، ليست هناك حاجة لتغيير أوزان الروابط. حيث تنتج شبكة الـ *perceptron* بمتجه الوزن $(0, 1.8)$ المخرجات المستهدفة لجميع سجلات البيانات في مجموعة البيانات التدريبية حيث يتم الانتهاء من تعلم أوزان الروابط لسجلات البيانات في مجموعة البيانات التدريبية بعد التكرار الأول لتغيير أوزان الروابط مع متجه الوزن النهائي $(0, 1.8)$. حد القرار هو المستقيم، $x_2 = 0$.

وبالنظر إلى المعادلات العامة لطريقة التعلم الخاص بتحديد أوزان الروابط:

$$\Delta w_j = \alpha(t_j - o_j)x = \alpha e_j x \qquad (١٣\text{-}٥)$$

$$w_j(k+1) = w_j(k) + \Delta w_j \quad (١٤-٥)$$

أو

$$\Delta w_{j,i} = \alpha(t_j - o_j)x_i = \alpha e_j x_i \quad (١٥-٥)$$

$$w_{j,i}(k+1) = w_{j,i}(k) + \Delta w_{j,i} \quad (١٦-٥)$$

حيث إن:

$e_j = t_j - o_j$ تمثل خطأ المخرجات

α هو معدل التعلم الذي يأخذ قيمة تتراوح في النطاق $(0,1)$

في المعادلة ٥-١١، يتم وضع قيمة α عند $1/2$. حيث إن التحيز (b) لوحدة المعالجة j هو وزن الرابط من المدخلات $x_0 = 1$ إلى وحدة المعالجة، فإنه يمكن التعويض في المعادلتين ٥-١٥، و٥-١٦ لتغيير التحيز الخاص بوحدة المعالجة j على النحو التالي:

$$\Delta b_j = \alpha(t_j - o_j) \times x_0 = \alpha(t_j - o_j) \times 1 = \alpha e_j \quad (١٧-٥)$$

$$b_j(k+1) = b_j(k) + \Delta b_j. \quad (١٨-٥)$$

٥-٣-٥ عيوب الشبكة العصبية الصناعية ذات التغذية الأمامية الأحادية الطبقة (Limitation of a Perceptron):

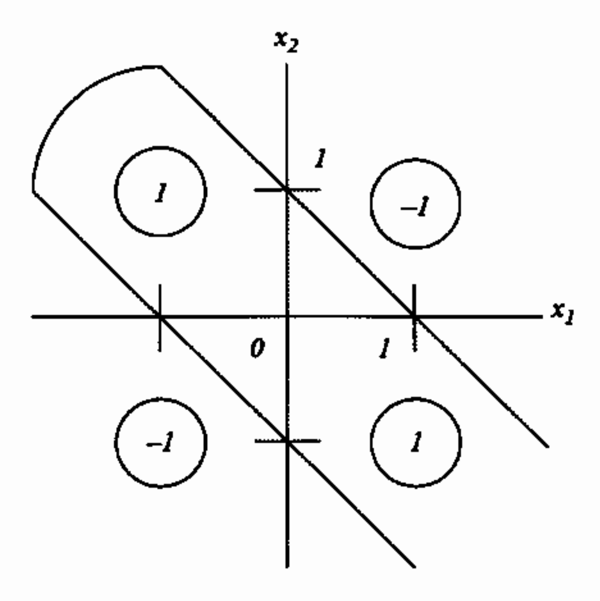
كما هو موضح في الأجزاء ٥-٣-٢ و٥-٣-٣، فإن كل وحدة من وحدات المعالجة تطبق حد القرار الخطي، وهو ما يعني دالة قابلة للفصل خطياً. حتى مع وجود وحدات معالجة متعددة في طبقة واحدة، تقتصر شبكة الـ *perceptron* على تطبيق دالة قابلة للفصل خطياً. على سبيل المثال، الدالة *XOR* في الجدول ٥-٣ ليست دالة قابلة للفصل خطياً. هناك مخرجة واحدة فقط للدالة *XOR* باستخدام وحدة معالجة واحدة لتمثيل المخرجات، يكون لدينا حد قرار واحد، وهو خط مستقيم يمثل دالة خطية.

على الرغم من ذلك، لا يوجد خط مستقيم في فضاء المدخلات لفصل نقطتي بيانات بها $o = 1$ عن نقطتي البيانات الأخريين التي بها مع $o = -1$. وهناك حاجة لحد قرار غير خطي، مثل ذلك الموضح في الشكل ٥-١٢ لفصل نقطتي البيانات التي بها $o = 1$ عن نقطتي البيانات الأخريين التي بها $o = -1$. لاستخدام وحدتي معالجة تطبق دوال قابلة للفصل خطياً لبناء شبكة ANN تطبيق الدالة XOR، فإننا نحتاج وحدات معالجة في تطبيق واحدة (الطبقة المخفية) لتطبيق حدي قرار، ووحدة معالجة واحدة في طبقة أخرى (طبقة المخرجات) للجمع بين مخرجات الوحدتين المخفيتين كما هو مبين في الجدول ٥-٤، والشكل ٥-٧. يعرف الجدول ٥-٥ دالة NOT المنطقية المستخدمة في الجدول ٥-٤. ومن ثم، نحتاج إلى شبكة ANN ثنائية الطبقات تطبق الدالة XOR، وهي دالة قابلة للفصل بشكل غير خطي.

يمكن استخدام طريقة التعلم الموصوفة من خلال المعادلات من ٥-١٣ إلى ٥-١٨، لمعرفة أوزان الروابط لكل وحدة من وحدات المخرجات باستخدام مجموعة من البيانات التدريبية، لأن القيمة المستهدفة t لكل وحدة من وحدات المخرجات تكون معطاة في البيانات التدريبية. وبالنسبة لكل وحدة مخفية، المعادلات من ٥-١٣، إلى ٥-١٨ هي معادلات غير قابلة للتطبيق لأننا لا نعرف قيمة t للوحدة المخفية. ومن ثم، فإننا نواجه صعوبة في معرفة أوزان الروابط والتحيز من البيانات التدريبية لشبكة الـ ANN المتعددة الطبقات. يتم التغلب على هذه الصعوبة لشبكات الـ ANN المتعددة الطبقات من خلال طريقة التعلم بالتوالد الخلفي كما سيتم توضيحه في الجزء التالي.

الشكل (٥-١٢)

نقاط البيانات الأربع للدالة *XOR*



الجدول (٥-٤)

دالة خاصة بكل وحدة معالجة في شبكة الـ *ANN* الثنائية الطبقات لتطبيق الدالة *XOR*

o_3= AND o_2	o_2=NOT(x_1 OR x_2)	o_1=x_1 OR x_2	x_2	x_1
-1	1	-1	-1	-1
1	1	1	1	-1
1	1	1	-1	1
-1	-1	1	1	1

الجدول (٥-٥)

الدالة *NOT*

o	x
1	-1
-1	1

٥-٤ طريقة التعلم بالتوالد الخلفي للشبكات العصبية الصناعية ذات التغذية الأمامية متعددة الطبقات

(Back-Propagation Learning Method for a Multilayer Feedforward ANN):

تهدف طريقة التعلم بالتوالد الخلفي (*back propagation learning method*) للشبكات العصبية الصناعية *ANN* ذات التغذية الأمامية متعددة الطبقات (*Rumelhart et al., 1986*) إلى البحث عن مجموعة من أوزان الروابط (بما في ذلك التحيزات) *W* التي تقلل من خطأ المخرجات. يتم تعريف خطأ المخرجات لسجل بيانات تدريبية *d* على النحو التالي:

$$E_d(W) = \frac{1}{2}\sum_j \left(t_{j,d} - o_{j,d}\right)^2 \qquad (١٩-٥)$$

حيث إن:

$t_{j,d}$ هي المخرجات المستهدفة لوحدة المخرجات *j* لسجل بيانات تدريبية *d*

$o_{j,d}$ هي المخرجات الفعلية التي تنتجها وحدة المخرجات *j* في شبكة الـ *ANN* المحتوية على الأوزان *W* لسجل البيانات التدريبية *d*

يتم تعريف خطأ المخرجات لمجموعة سجلات بيانات تدريبية على النحو التالي:

$$E(W) = \frac{1}{2}\sum_d \sum_j \left(t_{j,d} - o_{j,d}\right)^2. \qquad (٢٠-٥)$$

لأن كل $o_{j,d}$ تعتمد على *W*، فإن *E* هي دالة من *W*. تبحث طريقة التعلم بالتوالد الخلفي في فضاء الأوزان الممكنة، وتقيم مجموعة معطاة من الأوزان على أساس قيم *E* المرتبطة بها. وتسمى عملية البحث هذه بالبحث الهابط المتدرج (*gradient descent*

(*search*) الذي يغير الأوزان عن طريق تحريكهم في اتجاه تقليل خطأ المخرجات بعد اجتياز مدخلات سجل البيانات d من خلال شبكة الـ *ANN* بالأوزان *W*، على النحو التالي:

$$\Delta w_{j,i} = -\alpha \frac{\partial E_d}{\partial w_{j,i}} = -\alpha \frac{\partial E_d}{\partial net_j}\frac{\partial net_j}{\partial w_{j,i}} = \alpha \delta_j \frac{\partial\left(\sum_k w_{j,k}\tilde{o}_k\right)}{\partial w_{j,i}} = \alpha \delta_j \tilde{o}_i \qquad (٥-٢١)$$

حيث يتم تعريف δ_j على أنها:

$$\delta_j = -\frac{\partial E_d}{\partial net_j}\ , \qquad (٥-٢٢)$$

حيث إن:

α هو معدل التعلم بقيمة عادة تكون في النطاق $(0,1)$

$\tilde{o}_i$ هي المدخلات i إلى وحدة المعالجة j

إذا كانت الوحدة j تستقبل مباشرة مدخلات الشبكة الـ *ANN*؛ فإن $\tilde{o}_i$ هي x_i وخلاف ذلك، فإن $\tilde{o}_i$ هي من وحدة في الطبقة السابقة التي تغذي مخرجاتها كمدخلات إلى الوحدة j، لتغيير التحيز الخاص بوحدة المعالجة، يتم تعديل المعادلة ٥-٢١ باستخدام $\tilde{o}_i = 1$ على النحو التالي:

$$\Delta b_j = \alpha \delta_j \qquad (٥-٢٣)$$

إذا كانت الوحدة j هي وحدة مخرجات،

$$\delta_j = -\frac{\partial E_d}{\partial net_j} = -\frac{\partial E_d}{\partial o_j}\frac{\partial o_j}{\partial net_j} = -\frac{\partial\left(\frac{1}{2}\sum_j (t_{j,d} - o_{j,d})^2\right)}{\partial o_j}\frac{\partial\left(f_j(net_j)\right)}{\partial net_j} = \left(t_{j,d} - o_{j,d}\right) f_j'(net_j), \qquad (٥-٢٤)$$

حيث تدل f' على مشتق الدالة f فيما يتعلق بـ net. للحصول على قيمة للحد $f'_j(net_j)$ في المعادلة ٥-٢٤، يجب أن تكون دالة التحويل f للوحدة j شبه خطية، غير تنازلية، وقابلة للتفاضل، على سبيل المثال، خطية، سينية، وتماسية. بالنسبة لدالة التحويل السينية:

$$o_j = f_j(net_j) = \frac{1}{1+e^{-net_j}} \ ,$$

يكون لدينا ما يلي:

$$f'_j(net_j) = \frac{1}{1+e^{-net_j}} \frac{e^{-net_j}}{1+e^{-net_j}} = o_j(1-o_j). \qquad (٥-٢٥)$$

إذا كانت الوحدة j هي وحدة مخفية تقوم بتغذية مخرجاتها كمدخلات لوحدات المخرجات،

$$\delta_j = -\frac{\partial E_d}{\partial net_j} = -\frac{\partial E_d}{\partial o_j}\frac{\partial o_j}{\partial net_j} = -\frac{\partial E_d}{\partial o_j} f'_j(net_j) = -\left(\sum_n \frac{\partial E_d}{\partial net_n}\frac{\partial net_n}{\partial o_j}\right) f'_j(net_j),$$

حيث net_n هو المجموع الصافي لوحدة المخرجات n. باستخدام المعادلة ٥-٢٢، نعيد كتابة δ_j على النحو التالي:

$$\delta_j = \left(\sum_n \delta_n \frac{\partial net_n}{\partial o_j}\right) f'_j(net_j) = \left(\sum_n \delta_n \frac{\partial(\sum_j w_{n,j} o_j)}{\partial o_j}\right) f'_j(net_j) \qquad (٥-٢٦)$$
$$= \left(\sum_n \delta_n w_{n,j}\right) f'_j(net_j).$$

حيث إننا نحتاج δ_n في المعادلة ٥-٢٦، والتي يتم حسابها لوحدة المخرجات n، فإن تغيير الأوزان الخاصة بشبكة الـ ANN يجب أن تبدأ بتغيير أوزان وحدات المخرجات، والانتقال إلى تغيير الأوزان للوحدات المخفية في الطبقة السابقة بحيث إن δ_n لوحدة المخرجات n

يمكن استخدامها في حساب δ_j للوحدة المخفية j، وبعبارة أخرى، δ_n لوحدة المخرجات n يتم توالدها خلفياً لحساب δ_i للوحدة المخفية j، والتي يُطلق عليها التعلم بالتوالد الخلفي. التغييرات الخاصة بالأوزان والتحيزات، على النحو الذي تحدده المعادلات ٥-٢١ و٥-٢٣، يتم استخدامها لتحديث الأوزان والتحيزات للشبكة العصبية الصناعية ANN على النحو التالي:

$$w_{j,i}(k+1) = w_{j,i}(k) + \Delta w_{j,i} \quad (٥\text{-}٢٧)$$

$$b_j(k+1) = b_j(k) + \Delta b_j. \quad (٥\text{-}٢٨)$$

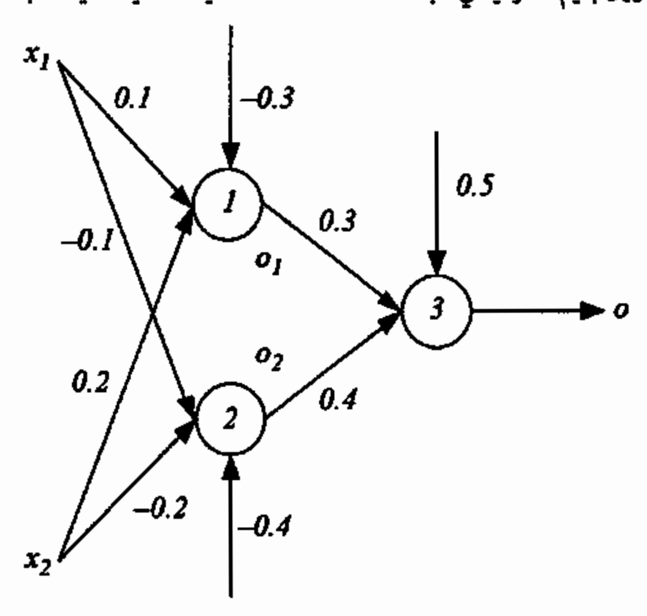
المثال (٥-٢)

ليكن لدينا شبكة ANN تستخدم دالة XOR وسجل البيانات الأول في الجدول ٥-٣ بحيث تكون $x_1 = -1$ ،$t = -1$ $x_2 = -1$ ، قم باستخدام طريقة التوالد الخلفي لتحديث الأوزان والتحيزات الخاصة بشبكة ANN في شبكة الـ ANN، يتم استخدام دالة التحويل السينية من قبل كل من الوحدتين المخفيتين، والدالة الخطية من قبل وحدة المخرجات. تبدأ شبكة الـ ANN بالقيم العشوائية التالية للأوزان والتحيزات في $(-1,1)$ كما هو مبين في الشكل ٥-١٣:

$$w_{1,1} = 0.1 \quad w_{2,1} = -0.1 \quad w_{1,2} = 0.2 \quad w_{2,2} = -0.2 \quad b_1 = -0.3$$
$$b_2 = -0.4 \quad w_{3,1} = 0.3 \quad w_{3,2} = 0.4 \quad b_3 = 0.5.$$

الشكل (٥-١٣)

مجموعة من الأوزان بقيم عشوائية في شبكة الـ *ANN* ذات التغذية الأمامية ثنائية الطبقات للدالة ***XOR***



قم باستخدام معدل التعلم *a=0.3*. بتمرير مدخلات سجل البيانات، $x_1 = -1$ ، و $x_2 = -1$ ، من خلال شبكة الـ *ANN* نحصل على ما يلي:

$$o_1 = sig(w_{1,1}x_1 + w_{1,2}x_2 + b_1) = sig(0.1 \times (-1) + 0.2 \times (-1) + (-0.3))$$
$$= sig(-0.6) = \frac{1}{1 + e^{-(-0.6)}} = 0.3543$$

$$o_2 = sig(w_{2,1}x_1 + w_{2,2}x_2 + b_2) = sig((-0.1) \times (-1) + (-0.2) \times (-1) + (-0.4))$$
$$= sig(-0.2) = \frac{1}{1 + e^{-(-0.2)}} = 0.4502$$

$$o = sig(w_{3,1}o_1 + w_{3,2}o_2 + b_3) = sig(0.3 \times 0.3543 + 0.4 \times 0.4502 + 0.5)$$
$$= sig(0.7864) = \frac{1}{1 + e^{-0.7864}} = 0.6871$$

بما أن الفرق بين $o=0.6871$ و $t= -1$ كبير، نحتاج إلى تغيير الأوزان والتحيزات بشبكة الـ ANN تُستخدم المعادلات ٥-٢١ و٥-٢٣ لتحديد التغييرات في الأوزان والتحيزات الخاصة بوحدة المخرجات كما يلي:

$$\Delta w_{3,1} = \alpha\delta_3\tilde{o}_1 = 0.3 \times \delta_3 \times o_1 = 0.3 \times \delta_3 \times 0.3543$$
$$\Delta w_{3,2} = \alpha\delta_3\tilde{o}_1 = 0.3 \times \delta_3 \times o_2 = 0.3 \times \delta_3 \times 0.4502$$
$$\Delta b_3 = \alpha\delta_3 = 0.3 \times \ \delta_3$$

وتُستخدَم المعادلة ٥-٢٤ لإيجاد δ_3، ثم تُستخدم δ_3 لإيجاد $\Delta w_{3,1}$،$\Delta w_{3,2}$ وΔb_3 على النحو التالي :

$$\delta_3 = (t - o)\ f_3'(net_3) = (t_{j,d} - o_{j,d})lin'(net_3) = (-1 - 0.681) \times 1 = -1.6871$$
$$\Delta w_{3,1} = 0.3 \times \delta_3 \times 0.3543 = 0.3 \times (-1.6871) \times 0.3543 = -0.1793$$
$$\Delta w_{3,2} = 0.3 \times \delta_3 \times 0.4502 = 0.3 \times (-1.6871) \times 0.4502 = -0.2279$$
$$\Delta b_3 = 0.3 \times \ \delta_3 = 0.3 \times (-1.6871) = -0.5061$$

تُستخدَم المعادلات ٥-٢١، ٥-٢٣، ٥-٢٥، ٥-٢٦، لتحديد التغييرات في الأوزان والتحيزات لكل وحدة مخفية على النحو التالي:

$$\delta_1 = \left(\sum_n \delta_n w_{n,1}\right) f_1'(net_1) = (\sum_{n=3}^{n=3} \delta_n w_{n,1}) f_1'(net_1)$$
$$= \delta_3 w_{3,1} o_1 (1 - o_1) = (-1.6871) \times 0.3 \times 0.3543 \times (1 - 0.3543) = -0.0471$$

$$\delta_2 = \left(\sum_n \delta_n w_{n,2}\right) f_2'(net_2) = (\sum_{n=3}^{n=3} \delta_n w_{n,2}) f_2'(net_2) = \delta_3 w_{3,2} o_2 (1 - o_2)$$

$$= (-1.6871) \times 0.4 \times 0.4502 \times (1 - 0.4502) = -0.0510$$

$$\Delta w_{1,1} = \alpha\delta_1 x_1 = 0.3 \times \delta_1 \times x_1 = 0.3 \times (-0.0471) \times (-1) = 0.0141$$

$$\Delta w_{1,2} = \alpha\delta_1 x_2 = 0.3 \times \delta_1 \times x_2 = 0.3 \times (-0.0471) \times (-1) = 0.0141$$

$$\Delta w_{2,1} = \alpha\delta_2 x_1 = 0.3 \times \delta_2 \times x_1 = 0.3 \times (-0.0510) \times (-1) = 0.0153$$

$$\Delta w_{2,2} = \alpha\delta_2 x_2 = 0.3 \times \delta_2 \times x_2 = 0.3 \times (-0.0510) \times (-1) = 0.0153$$

$$\Delta b_1 = \alpha\delta_1 = 0.3 \times (-0.0471) = -0.0141$$

$$\Delta b_2 = \alpha\delta_2 = 0.3 \times (-0.0510) = -0.0153.$$

باستخدام التغييرات على جميع الأوزان والتحيزات الخاصة بشبكة الـ *ANN*، تُستخدَم المعادلات ٥-٢٧ و٥-٢٨ لتنفيذ التكرار الخاص بتحديث الأوزان والتحيزات على النحو التالي:

$$w_{1,1}(1) = w_{1,1}(0) + \Delta w_{1,1} = 0.1 + 0.0141 = 0.1141$$

$$w_{1,2}(1) = w_{1,2}(0) + \Delta w_{1,2} = 0.2 + 0.0141 = 0.2141$$

$$w_{2,1}(1) = w_{2,1}(0) + \Delta w_{2,1} = -0.1 + 0.0153 = -0.0847$$

$$w_{2,2}(1) = w_{2,2}(0) + \Delta w_{2,2} = -0.2 + 0.0153 = -0.1847$$

$$w_{3,1}(1) = w_{3,1}(0) + \Delta w_{3,1} = 0.3 - 0.1793 = 0.1207$$

$$w_{3,2}(1) = w_{3,2}(0) + \Delta w_{3,2} = 0.4 - 0.2279 = 0.1721$$

$$b_1(1) = b_1(0) + \Delta b_1 = -0.3 - 0.0141 = -0.3141$$

$$b_2(1) = b_2(0) + \Delta b_2 = -0.4 - 0.0153 = -0.4153$$

$$b_3(1) = b_3(0) + \Delta b_3 = 0.5 - 0.5061 = -0.0061$$

سيتم استخدام هذه المجموعة الجديدة للأوزان والتحيزات، $w_{j,i}(1)$، و$b_j(1)$ لتمرير مدخلات سجل البيانات الثاني من خلال شبكة الـ *ANN*، ومن ثمّ تحديث الأوزان والتحيزات مرة أخرى للحصول على $w_{j,i}(2)$، و$b_J(2)$ إذا لزم الأمر. تتكرر هذه العملية مرةً أخرى لسجل البيانات الثالث، وسجل البيانات الرابع، والعودة إلى سجل البيانات الأول، وهلُم جرا،

حتى يصبح مقياس خطأ المخرجات E على النحو المحدد في المعادلة ٥-٢٠ أصغر من الحد المحدد مسبقاً، على سبيل المثال، القيمة: 0.1.

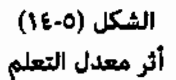
يمكن استخدام مقياس خطأ المخرجات، مثل E، أو خطأ متوسط الجذر التربيعي (*root-mean-square error*) على كافة سجلات البيانات التدريبية ليحدد متى يتوقف تعلم الأوزان والتحيزات الخاصة بشبكة *ANN*. عدد مرات التكرار، على سبيل المثال ١٠٠٠ تكرار، هو معيار آخر والذي يمكن استخدامه لوقف التعلم.

يُسمى تحديث الأوزان والتحيزات بعد تمرير كل سجل من سجلات البيانات في مجموعة البيانات التدريبية بالتعلم المتزايد (*Incremental learning*). في التعلم المتزايد، يتم تحديث الأوزان والتحيزات بحيث إنها سوف تعمل على نحو أفضل لسجل بيانات واحد. التغييرات القائمة على سجل بيانات واحد قد تذهب في اتجاه مختلف، بحيث تتلاشى التغييرات التي تمّ إجراؤها لسجل بيانات آخر، مما يجعل عملية التعلم تستغرق وقتاً طويلاً لتتقارب إلى المجموعة النهائية للأوزان والتحيزات التي تتناسب لكل سجلات البيانات. التعلم بالدفعة الواحدة *(batch learning)* ينبغي أن يوقف تحديث الأوزان والتحيزات حتى يتم تمرير كافة سجلات البيانات في مجموعة البيانات التدريبية من خلال شبكة الـ *ANN*، وحتى يتم احتساب كل التغييرات المرتبطة بالأوزان والتحيزات وحساب متوسطاتها. يتغير متوسط الوزن والتحيز لجميع سجلات البيانات، وهو ما يعني، أنه يتم استخدام الأثر الكلي للتغيرات على الأوزان والتحيزات من قبل جميع سجلات البيانات، بغرض تحديث الأوزان والتحيزات.

يؤثر معدل التعلم (*learning rate*) أيضاً على جودة وسرعة استمرار التعلم. كما هو موضح في الشكل ٥-١٤، فإن معدل التعلم بقيمة صغيرة، على سبيل المثال 0.01 ينتج عنه تغيير صغير للأوزان والتحيزات، ومن ثم يكون هناك انخفاض طفيف في E، ويجعل عملية التعلم تستغرق وقتاً طويلاً للوصول إلى قيمة الحد الأدنى العامة لـ E أو قيمة الحد الأدنى المحلية لـ E. على الجانب الآخر، ينتج معدل التعلم ذو القيمة الكبيرة تغييراً كبيراً في الأوزان والتحيزات، الأمر الذي قد يسبب في أن عملية البحث عن W لتقليل قيمة E لا تصل إلى قيمة الحد الأدنى المحلية أو العامة لـ E. ومن هنا، في مفاضلة بين معدل التعلم ذي القيمة الصغيرة ومعدل التعلم ذي القيمة الكبيرة، يمكن استخدام طريقة معدلات التعلم المتكيفة

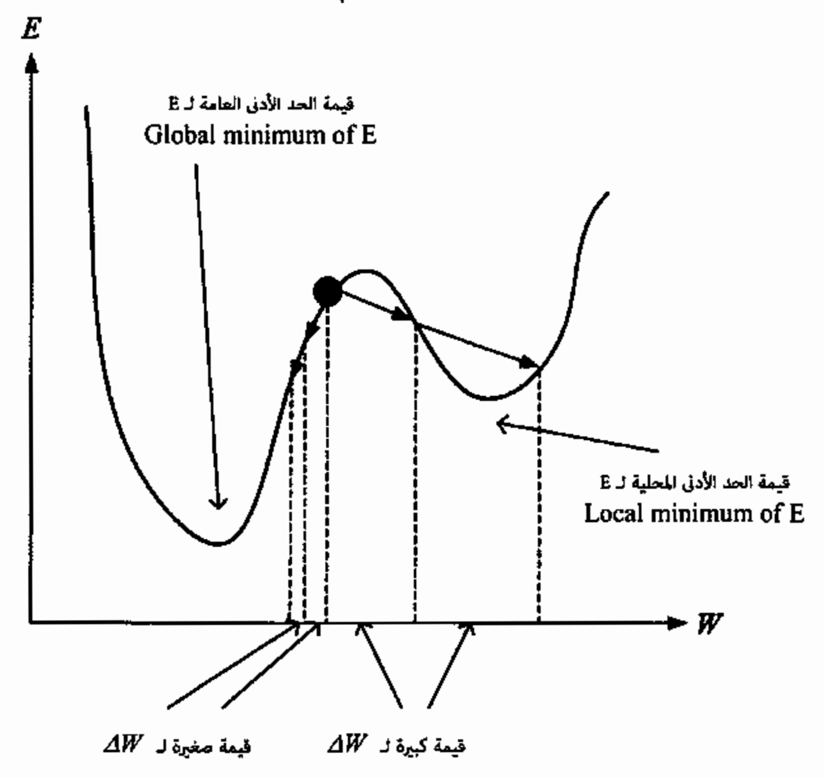
بحيث تبدأ بمعدل تعلم كبير لتسريع عملية التعلم، ثم القيام بالتغيير إلى معدل تعلم صغير لأخذ خطوات صغيرة للوصول إلى قيمة الحد الأدنى لـ E المحلية أو العامة.

على عكس أشجار القرار في الفصل ٤، لا تُظهر أي شبكة عصبية صناعية ANN نموذجاً واضحاً وصريحاً للتصنيف ودالة تنبؤ تتعلمها شبكة الـ ANN من خلال البيانات التدريبية. يتم تمثيل الدالة ضمنياً من خلال أوزان الروابط، والتحيزات والتي لا يمكن ترجمتها إلى أنماط تصنيف وتنبؤ ذات معنى في نطاق المشكلة المبحوثة. على الرغم من أن المعرفة بأنماط التصنيف والتنبؤ قد تمّ الحصول عليها من خلال شبكة الـ ANN، فإن هذه المعرفة غير متوفرة بشكل قابل للتفسير. ومن ثم، تساعد الشبكات العصبية الصناعية على أداء مهمة التصنيف والتنبؤ، وليس على أداء مهمة اكتشاف المعرفة.



الشكل (٥-١٤)

أثر معدل التعلم



٥-٥ الاختيار التجريبي لمعمارية الشبكة العصبية الصناعية من أجل ملائمة جيدة للبيانات (Empirical Selection of an ANN Architecture for a Good Fit to Data):

على عكس نماذج الانحدار في الفصل ٢، لا تتطلب دالة تعلم التصنيف والتنبؤ من خلال الشبكة العصبية الصناعية ذات التغذية الأمامية المتعددة الطبقات *ANN* تعريف شكل معين لتلك الدالة، مما يجعل الأمر صعباً عندما تكون مجموعة البيانات كبيرة، ونحن لدينا معرفة مسبقة قليلة عن المجال أو البيانات. تعتمد كثيراً درجة تعقيد الشبكة العصبية الصناعية *ANN* والدالة التي تتعلمها وتمثلها شبكة الـ *ANN* على عدد الوحدات المخفية. فكلما زادت الوحدات المخفية لدى شبكة الـ *ANN* ، أصبحت الدالة التي تتعلمها وتمثلها شبكة الـ *ANN* أكثر تعقيداً ولكن، إذا كان لنا أن نستخدم شبكة الـ *ANN* معقدة لتعلم دالة بسيطة، فإننا قد نرى دالة شبكة الـ *ANN* مفرطة في مطابقة البيانات (*data over-fitting*)، ومن ثم تكون غير مناسبة، كما هو موضح في الشكل ٥-١٥. في هذا الشكل، يتم توليد نقاط البيانات باستخدام النموذج الخطي:

$$y = x + \varepsilon,$$

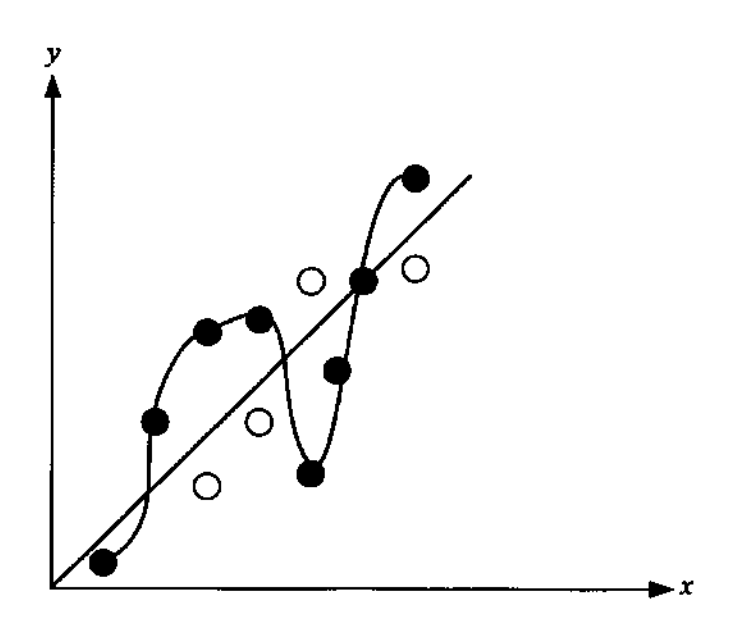
حيث يدل الرمز ε على الخطأ العشوائي. ومع ذلك، تمّ تركيب نموذج غير خطي لنقاط البيانات التدريبية كما هو موضح بالدوائر الداكنة في الشكل ٥-١٥، والتي تغطي كل نقطة بيانات تدريبية مع عدم وجود فرق بين القيمة الهدف لـ y والقيمة المتوقعة لـ y من النموذج غير الخطي. على الرغم من أن النموذج غير الخطي يوفر حلاً مثالياً للبيانات التدريبية، إلا أن الأداء التنبؤي للنموذج غير الخطي على لنقاط بيانات جديدة في مجموعة البيانات الاختبارية كما هو موضح بالدوائر البيضاء في الشكل ٥-١٥ سيكون أكثر سوءاً من تلك الموجودة بالنموذج الخطي، $y = x$، وذلك للأسباب التالية:

- يلتقط النموذج غير الخطي الخطأ العشوائي ε في النموذج.
- إن الأخطاء العشوائية لنقاط بيانات جديدة تتصرف بشكل مستقل، ومختلف عن الأخطاء العشوائية لنقاط البيانات التدريبية.
- إن الأخطاء العشوائية لنقاط البيانات التدريبية التي يتم التقاطها في النموذج غير الخطي لا تتطابق تماماً مع الأخطاء العشوائية لنقاط البيانات الجديدة في مجموعة البيانات الاختبارية، مما يسبب أخطاء في التنبؤ.

وبشكل عام، فإن أي نموذج مفرط في المطابقة لا يتم تعميمه بشكل جيد لنقاط بيانات جديدة في مجموعة البيانات الاختبارية. عندما لا يكون لدينا معرفة مسبقة بمجموعة بيانات معينة (على سبيل المثال، الشكل أو تعقيد دالة التصنيف والتنبؤ)، ينبغي علينا القيام بالمحاولة تجريبياً لعمل معماريات لشبكة الـ ANN بمستويات متفاوتة من التعقيد باستخدام أعداد مختلفة من الوحدات المخفية. يتم تدريب كل معمارية لشبكة الـ ANN لتعلُّم أوزان وتحيزات الروابط في مجموعة البيانات التدريبية، ويتم اختبار أدائها التنبؤي على مجموعة بيانات اختبارية. يتم اعتبار معمارية شبكة الـ ANN ذات الأداء الجيد على البيانات الاختبارية أنها تعطي تطابقاً وملاءمةً جيدةً للبيانات ومن ثم يتم اختيارها.

الشكل (٥-١٥)

مثال يوضح نموذجاً غير خطي مفرط في مطابقة البيانات من نموذج خطي



٥-٦ البرمجيات والتطبيقات (Software and Applications):

يحتوي الموقع الإلكتروني (*http://www.knuggets.com*) على معلومات عن أدوات استكشاف بيانات متنوعة. توفر حزم البرمجيات التالية أدوات برمجية للشبكات العصبية الصناعية *ANNs* باستخدام طريقة التعلم بالتوالد الخلفي:

- *Weka (http://www.cs.waikato.ac.nz/ml/weka/)*
- *MATLAB® (www.mathworks.com/)*

بعض التطبيقات الخاصة بالشبكات العصبية الصناعية *ANNs* يمكن العثور عليها في: (*Ye et al., 1993; Ye, 1996, 2003, Chapter 3; Ye and Zhao, 1996, 1997*).

التمارين (Exercises) :

٥-١ مجموعة البيانات التدريبية للدالة المنطقية، $y= NOT\ x$، معطاة في الجدول المرفق. استخدام الطريقة البيانية لتحديد حد القرار، والوزن، والتحيز للشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة *(perceptron)* أحادية الوحدة لهذه الدالة المنطقية.

مجموعة البيانات التدريبية:

X	Y
-1	1
1	-1

٥-٢ بالنظر في الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة *(perceptron)* أحادية الوحدة في التمرين ٥-١. أسند القيمة *0.2* كقيمة أولية للأوزان والتحيز واستخدام معدل التعلم *0.3*. استخدام طريقة التعلم لعمل تكرار واحد لتحديث الوزن والتحيز لسجلي البيانات الاثنين للدالة المنطقية في التمرين ٥-١.

٥-٣ مجموعة البيانات التدريبية لدالة تصنيف ذات ثلاثة متغيرات خاصية ومتغير هدف واحد معطاة أدناه. استخدام الطريقة البيانية لتحديد حد القرار، والوزن، والتحيز للشبكة العصبية الصناعية ذات التغذية الأمامية الأحادية الطبقة *(perceptron)* أحادية الوحدة لدالة التصنيف تلك.

مجموعة البيانات التدريبية:

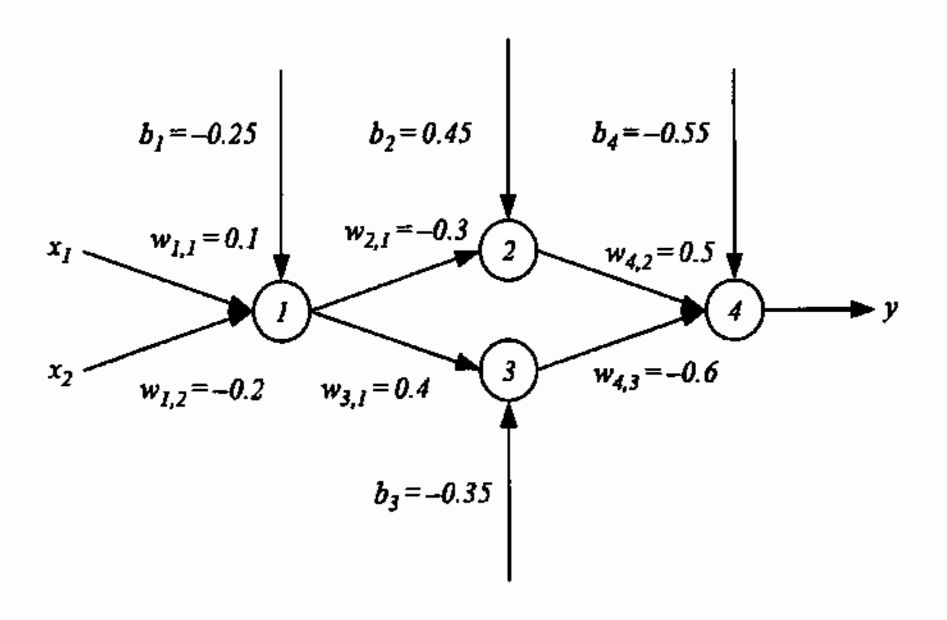
x_1	x_2	x_3	y
-1	-1	-1	-1
-1	-1	1	-1
-1	1	-1	-1
-1	1	1	1
1	-1	-1	-1
1	-1	1	1
1	1	-1	1
1	1	1	1

٥-٤ تُستخدَم الشبكة العصبية الصناعية ذات التغذية الأمامية أحادية الطبقة *(perceptron)* أحادية الوحدة لتعلم دالة التصنيف في التمرين ٥-٣. أسند القيمة *0.4* كقيمة أولية للأوزان والتحيز واستخدام معدل التعلم *0.2*. استخدام طريقة التعلم لعمل تكرار واحد لتحديث الوزن والتحيز لسجلي البيانات الثالث والرابع لهذه الدالة.

٥-٥ لنفترض أن لدينا شبكة عصبية صناعية ذات تغذية أمامية ثنائية الطبقة ومترابطة ترابطاً كاملاً بمتغير مدخلات واحد، ووحدة واحدة مخفية، ومتغيري مخرجات اثنين. أسند القيمة *0.1* كقيمة أولية. للأوزان والتحيزات، واستخدام معدل التعلم *0.3*. دالة التحويل المستخدمة *(sigmoid function)* هي الدالة السينية لكل وحدة. قم بإظهار التصميم الخاص بالشبكة العصبية الصناعية *ANN* وقم بعمل تكرار واحد لتحديث الوزن والتحيز باستخدام خوارزمية التعلم بالتوالد الخلفي، والمثال التدريبي التالي:

x	y_1	y_2
1	0	1

٦-٥ تُستخدَم شبكة العصبية الصناعية الـ *ANN* التالية مع ذات الأوزان والتحيز المبدئي لتعلم دالة *XOR* دالة التحول للوحدات ١ و ٤ هي الدالة الخطية. دالة التحول للوحدات ٢ و ٣ هي دالة التحول السينية. معدل التعلم هو$a=0.3$. أعمل تكرار واحد لتحديث الوزن والتحيز لـ $w1, 1$، $w1, 2$، $w2, 1$، $w3, 1$، $w4, 2$، $w4,3$، $b2$ بعد تغذية المتغيرات بالقيم $x1 = 0$ ، $x2=1$ في شبكة الـ *ANN* .



:*XOR*

y	x_1	x_1
0	0	0
1	1	0
1	0	1
0	1	1

٦- الدعم الآلي المتجه
Support Vector Machines

يقوم الدعم الآلي المتجه (*Support Vector Machines-SVM*) بتعريف دالة بفئتين مستهدفتين (*two target classes*) من خلال حل مسألة برمجية تربيعية *(quadratic programming problem)*. في هذا الفصل، نستعرض بإيجاز الأساس النظري للدعم الآلي المتجه (*SVM*) الذي يؤدي إلى صياغة مسألة برمجية تربيعية لتعلُّم مصنِّف ما. نقوم بعد ذلك باستعراض صياغة الدعم الآلي المتجه (*SVM*) لمصنِّف خطي *(linear classifier)*، ولمسألة قابلة للانفصال خطياً *(linearly separable problem)*، تليها صياغة الدعم الآلي المتجه (*SVM*) لمصنِّف خطي ولمسألة قابلة للانفصال بشكل غير خطي، وصياغة الدعم الآلي المتجه (*SVM*) لمصنِّف غير خطي ولمسألة قابلة للانفصال بشكل غير خطي باستخدام دوال كيرنل *(kernel functions)*. نقوم أيضاً باستعراض طرق لتطبيق الدعم الآلي المتجه (*SVM*) لدالة تصنيف بأكثر من فئتين مستهدفتين. وترد قائمة من حزم البرمجيات لغرض استكشاف البيانات تساند الدعم الآلي المتجه (*SVM*). وسيتم استعراض بعض التطبيقات الخاصة بالدعم الآلي المتجه (*SVM*) مع مراجعها.

٦-١ الأساس النظري لصياغة وحل مشكلة التحسين لتعلُّم دالة التصنيف
(Theoretical Foundation for Formulating and Solving an Optimization Problem to Learn a Classification Function):

بالنظر إلى مجموعة بها عدد n من نقاط البيانات $(x_1, y_1), \dots, (x_n, y_n)$، وإلى دالة تصنيف تطابق وتناسب البيانات، $y=f_A(x)$، حيث تأخذ y واحدة من القيم النوعية $\{-1,1\}$، و x هو متجه من المتغيرات ذو عدد p من الأبعاد، و A هو مجموعة من المعلمات *(parameters)* في الدالة f التي يتم تعلمها وتحديدها باستخدام البيانات التدريبية. على سبيل المثال، إذا تمّ استخدام الشبكة العصبية الصناعية *(ANN)* لتعريف وتمثيل دالة التصنيف f، فتكون أوزان الروابط والتحيزات هي المعلمات في f. تقوم مخاطر التصنيف المتوقعة *(the expected risk of classification)* باستخدام f لقياس خطأ التصنيف، وتُعرف بأنها:

$$R(A) = \int |f_A(x) - y|\, P(x,y)dxdy, \qquad (١-٦)$$

حيث تشير $P\ (x,y)$ إلى دالة الاحتمال لـ x و y. وتعتمد مخاطر التصنيف المتوقعة على قيم A. تشير القيمة الأقل لمخاطر التصنيف المتوقعة إلى أداء تعميم أفضل لدالة التصنيف، وذلك يعني أن تصبح دالة التصنيف قادرة على تصنيف المزيد من نقاط البيانات بشكل صحيح. المجموعات المختلفة من قيم A تعطي دوال تصنيف مختلفة $f_A(x)$، ومن ثم تنتج أخطاء تصنيف مختلفة ومستويات مختلفة من المخاطر المتوقعة. يتم تعريف المخاطر التجريبية على عينة من نقاط البيانات n كالتالي:

$$R_{emp}(A) = \frac{1}{n}\sum_{i=1}^{n} |f_A(x_i) - y_i|. \qquad (٢-٦)$$

يقدم فابنيك وتشيرفونينكيس (*Vapnik, 1989,2000*) القيد التالي على مخاطر التصنيف المتوقعة والذي يصبح نافذاً بالاحتمالية $1 - \eta$:

$$R(A) \le R_{emp}(A) + \sqrt{\frac{v\left(ln\frac{2n}{v} + 1\right) - ln\frac{\eta}{4}}{n}}, \qquad (٣-٦)$$

حيث يدل v على البعد الخاص بـ VC (*Vapnik and Chervonenkis*) لـ f_A ويقيس درجة تعقيد f_A، والذي يتم التحكم به بعدد المعلمات A في f للعديد من دوال التصنيف. ومن ثم، فإن مخاطر التصنيف المتوقعة تكون مقيدة بكلٍ من مخاطر التصنيف التجريبية، والحد الثاني في المعادلة ٦-٣ مع كون الحد الثاني يتزايد مع بُعد VC. لتقليل مخاطر التصنيف المتوقعة، نحتاج إلى تقليل كل من المخاطر التجريبية وبُعد VC لـ f_A في الوقت نفسه. وهذا ما يُسمى بمبدأ تقليل المخاطر الهيكلية. حيث إن تقليل قيمة بُعد VC لـ f_A، أو درجة تعقيد f_A، هو مثل البحث عن دالة تصنيف ذات طول وصف أدنى لعمل تعميم جيد كما تمّ مناقشته في الفصل ٤. يبحث الدعم الآلي المتجه (*SVM*) عن مجموعة من القيم A

والتي تقلل من المخاطر التجريبية، وعن قيمة بُعد *VC* في الوقت نفسه عن طريق صياغة وحل مشكلة التحسين أو المثالية *(Optimization problem)*، وتحديداً، مشكلة البرمجة التربيعية. توفر الأجزاء التالية صياغة الدعم الآلي المتجه *(SVM)* لمشكلة البرمجة التربيعية لثلاثة أنواع من مشاكل التصنيف: (١) المصنّف الخطي والمشكلة القابلة للانفصال خطياً، (٢) المصنّف الخطي والمشكلة القابلة للانفصال بشكل غير خطي، و(٣) المصنّف غير الخطي والمشكلة القابلة للانفصال بشكل غير خطي. وكما نوقش في الفصل ٥، فإن دالة *AND* المنطقية هي مشكلة تصنيف قابلة للانفصال خطياً، ولا تتطلب سوى المصنّف الخطي المذكور في النوع (١)، ودالة *XOR* المنطقية هي مشكلة تصنيف قابلة للانفصال بشكل غير خطي، والتي تتطلب المصنّف غير الخطي المذكور في النوع (٣). ولأن أي مصنف خطي عموماً يكون له قيمة أقل لبَعد *VC* أكثر من المصنّف غير الخطي، فإن استخدام المصنّف الخطي لمشكلة قابلة للانفصال بشكل غير خطي والمذكورة في النوع (٢) يمكن أن ينتج أحياناً حد أدنى لمخاطر التصنيف المتوقعة أقل من استخدام مصنّف غير خطي لمشكلة قابلة للانفصال بشكل غير خطي.

٦-٢ صياغة الدعم الآلي المتجه (SVM) لمصنف خطي ولمشكلة قابلة للانفصال خطياً (SVM Formulation for a Linear Classifier and a Linearly Separable Problem):

بالنظر في تعريف مصنّف خطي لشبكة عصبية صناعية ذات تغذية أمامية أحادية الطبقة *(perceptron)* في الفصل ٥:

$$f_{w,b}(x) = sign(w'x + b). \qquad (٦\text{-}٤)$$

حد القرار الذي يفصل فئتين مستهدفتين $\{-1,1\}$ هو:

$$w'x + b = 0. \qquad (٦\text{-}٥)$$

ويعمل المصنِّف الخطي بالطريقة التالية:

$$y = sign(w'x + b) = 1 \qquad \text{if } w'x + b > 0 \qquad (٦\text{-}٦)$$
$$y = sign(w'x + b) = -1 \quad \text{if } w'x + b \leq 0$$

إذا ما فرضنا القيد التالي:

$$\|w\| \leq M,$$

حيث إن M عبارة عن ثابت، وتدل $\|w\|$ على مقياس لمتجه w ذي عدد p من الأبعاد ويعرف أنه:

$$\|w\| = \sqrt{w_1^2 + \cdots + w_p^2}.$$

إن مجموعة الفضاءات الجزئية (*hyperplanes*) المعرفة كما يلي:

$$\{f_{w,b} = sign(w'x + b) | \|w\| \leq M\},$$

تحتوي على بُعد VC المسمى v الذي يحقق القيد (*Vapnik, 1989, 2000*):

$$v \leq \min\{M^2, p\} + 1. \qquad (٧\text{-}٦)$$

وبتخفيض قيمة $\|w\|$ ، ستنخفض قيمة M ومن ثم تنخفض قيمة البعد VC المسمى v. كما هو مطلوب من قبل مبدأ تقليل المخاطر الهيكلية تقليل المخاطر الهيكلية، نريد تخفيض قيمة $\|w\|$، أو ما يكافئها:

$$\min \frac{1}{2} \|w\|^2. \qquad (٨-٦)$$

تغيير قيمة w لا يغير ميل الفضاءات الجزئية لحد القرار. وتغيير قيمة b لا يغير ميل حد القرار، ولكنه يقوم بتحريك الفضاءات الجزئية لحد القرار بشكل متوازٍ. على سبيل المثال، في فضاء المتجه ثنائي الأبعاد كما هو مبين في الشكل رقم ٦-١، يكون حد القرار هو:

$$w_1 x_1 + w_2 x_2 + b = 0 \quad \text{or} \quad x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2}, \qquad (٩-٦)$$

ويكون ميل المستقيم لحد القرار هو $-w_1 / w_2$، وتكون نقطة التقاطع لمستقيم حد القرار هي $-b/ w_2$. إنّ تغيير قيمة w إلى القيمة $c_w w$ ، حيث c_w ثابت، لا يغير ميل مستقيم حد القرار لأن: $-c_w w_1/c_w w_2 = -w_1 / w_2$. وتغيير قيمة b إلى القيمة $c_b b$، حيث c_b هو ثابت، لا يغير أيضاً ميل مستقيم حد القرار، ولكنه يغير نقطة تقاطع المستقيم $-c_b b / w_2$ ومن ثم يتحرك الخط المستقيم بشكل متوازٍ.

ويبين الشكل ٦-١ أمثلة لنقاط بيانات بقيمة هدف تساوي 1 (يشار إليها بالدوائر الصغيرة)، وأمثلة لنقاط بيانات ذات القيمة الهدف 1- (المشار إليها بالمربعات الصغيرة). من بين نقاط البيانات بالقيمة الهدف المساوية 1، نأخذ في الاعتبار نقطة البيانات الأقرب إلى حد القرار، x_{+1} كما هو موضح بنقطة البيانات ذات الدائرة الداكنة في الشكل ٦-١. من بين نقاط البيانات بالقيمة الهدف المساوية لـ 1-، نأخذ في الاعتبار نقطة البيانات الأقرب إلى حد القرار، x_{-1} كما هو موضح بنقطة البيانات ذات المربع الداكن في الشكل ٦-١. لنفترض أنه بالنسبة للنقطتين x_{+1} وx_{-1} من نقاط البيانات يكون لدينا:

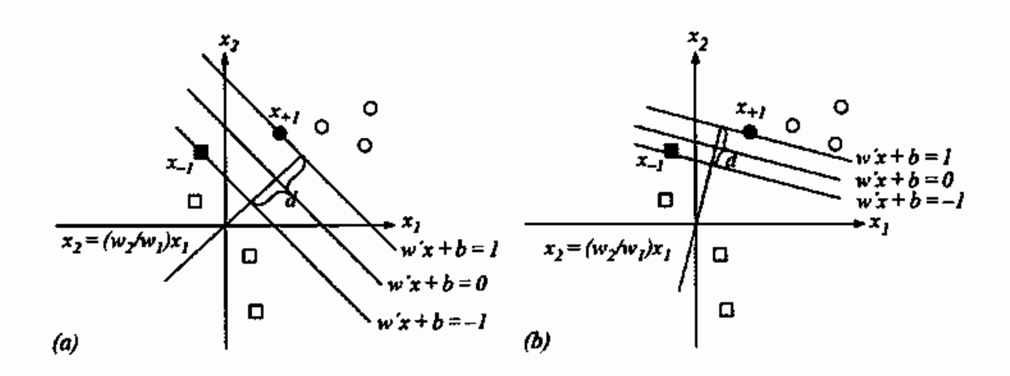
$$\begin{aligned} w' x_{+1} + b &= c_{+1} \\ w' x_{-1} + b &= c_{-1}. \end{aligned} \qquad (١٠-٦)$$

نريد تعديل قيمة w لتكون $c_w w$ وتعديل قيمة b لتكون $c_b b$ بحيث يكون لدينا:

$$c_w w' x_{+1} + c_b b = 1$$
$$c_w w' x_{-1} + c_b b = -1 \qquad (١١-٦)$$

الشكل (٦-١)

الدعم الآلي المتجه (SVM) لمصنِّف خطي ومشكلة قابلة للانفصال خطياً. (a) حد القرار ذو هامش كبير. (b) حد القرار ذو هامش صغير



ولا تزال تدل على القيم التي تمّ تغييرها بواسطة w و b. ويكون لدينا:

$$\min\{|w'x_i + b|, \quad i = 1, \ldots, n\} = 1,$$

وهو ما يعني ضمناً $|w' x + b|=1$ لنقطة البيانات في كل فئة مستهدفة أقرب إلى حد القرار $w'x+b=0$.

على سبيل المثال، في فضاء المتجه ثنائي الأبعاد x تصبح المعادلات ٦-١٠ و٦-١١ كما يلي:

$$w_1 x_{+1,1} + w_2 x_{+1,2} + b = c_{+1} \qquad (١٢-٦)$$

$$w_1 x_{-1,1} + w_2 x_{-1,2} + b = c_{-1} \quad (١٣-٦)$$

$$c_w w_1 x_{+1,1} + c_w w_2 x_{+1,2} + c_b b = 1 \quad (١٤-٦)$$

$$c_w w_1 x_{-1,1} + c_w w_2 x_{-1,2} + c_b b = -1. \quad (١٥-٦)$$

نقوم بحل المعادلات من ٦-١٢ إلى ٦-١٥ للحصول على c_w وc_b. علينا أولاً استخدام المعادلة ٦-١٤ للحصول على:

$$c_w = \frac{1 - c_b b}{w_1 x_{+1,1} + w_2 x_{+1,2}}, \quad (١٦-٦)$$

ونعوض عن c_w الموجودة في المعادلة ٦-١٦ داخل ٦-١٥ للحصول على:

$$\frac{1 - c_b b}{w_1 x_{+1,1} + w_2 x_{+1,2}}\left(w_1 x_{-1,1} + w_2 x_{-1,2}\right) + c_b b = -1. \quad (١٧-٦)$$

بعد ذلك نستخدم المعادلات ٦-١٢ و٦-١٣ للحصول على:

$$w_1 x_{+1,1} + w_2 x_{+1,2} = c_{+1} - b \quad (١٨-٦)$$

$$w_1 x_{-1,1} + w_2 x_{-1,2} = c_{-1} - b, \quad (١٩-٦)$$

ونعوض باستخدام المعادلات ٦-١٨ و٦-١٩ داخل المعادلة ٦-١٧ للحصول على:

$$\frac{1 - c_b b}{c_{+1} - b}(c_{-1} - b) + c_b b = -1$$

$$\frac{c_{-1}-b}{c_{+1}-b}-\frac{(c_{-1}-b)b}{c_{+1}-b}c_b+bc_b=-1$$

$$c_b=\frac{2b-c_{+1}-c_{-1}}{b^2+b-c_{-1}b}. \qquad (٦-٢٠)$$

وأخيراً، نستخدم المعادلة ٦-١٤ لحساب c_w، ونعوض بالمعادلات ٦-١٨ و٦-٢٠ في المعادلات الناتجة للحصول على:

$$c_w=\frac{1-c_bb}{w_1x_{+1,1}+w_2x_{+1,2}}=\frac{1-c_bb}{c_{+1}-b}=\frac{1-(2b-c_{+1}-c_{-1}/b+1-c_{-1})}{c_{+1}-b}$$

$$=\frac{1-b+c_{+1}}{(c_{+1}-b)(b+1-c_{-1})}. \qquad (٦-٢١)$$

المعادلات ٦-٢٠ و٦-٢١ توضح كيفية إعادة تقييم w وb في فضاء المتجه ثنائي الأبعاد x لتكن w و b تشير إلى القيم المتغيرة. الفضاء الجزئي يشطر (ينصّف) المستقيمين $w'x + b=1$ و$w'x + b=-1$ بالمستقيم $w'x + b=0$، كما هو مبين في الشكل ٦-١. أي نقطة x من نقاط البيانات ذات فئة مستهدفة 1+ تحقق:

$$w'x+b\geq 1$$

حيث إن نقطة البيانات ذات الفئة المستهدفة 1+ الأقرب إلى $w'x + b=0$ يكون لديها $w'x+ b=1$. أي نقطة x من نقاط البيانات ذات الفئة المستهدفة 1- تحقق:

$$w'x+b\leq -1$$

حيث إن نقطة البيانات ذات الفئة الهدف 1- الأقرب إلى $w'x + b=0$ يكون لديها $w'x +b= -1$ ومن ثم، فإن المصنّف الخطي يمكن تعريفه على النحو التالي:

$$y = sign(w'x + b) = 1 \qquad \text{if } w'x + b \geq 1 \qquad (٦-٢٢)$$
$$y = sign(w'x + b) = -1 \qquad \text{if } w'x + b \leq -1.$$

لتقليل قيمة المخاطر التجريبية R_{emp} أو خطأ التصنيف التجريبي كما هو مطلوب من مبدأ تقليل المخاطر الهيكلية المعرف بالمعادلة ٦-٣، فإننا نتطلب:

$$y_i(w'x_i + b) \geq 1, \quad i = 1, \ldots, n. \qquad (٦-٢٣)$$

إذا كانت $y_i=1$، فنحن نريد $w'x_i + b \geq 1$ بحيث ينتج المصنّف الخطي في المعادلة ٦-٢٢ الفئة المستهدفة 1. إذا كانت $y_i= -1$، فنحن نريد $w'x_i + b \leq -1$ بحيث ينتج المصنّف الخطي في المعادلة ٦-٢٢ الفئة المستهدفة 1-. ومن ثم، تحدد المعادلة ٦-٢٣ متطلبات التصنيف الصحيح لعينة من نقاط البيانات (x_i, y_i) ،$i=1, \ldots, n$. لذلك، بوضع المعادلات ٦-٨ و٦-٢٣ معاً يتيح لنا تطبيق مبدأ المخاطر الهيكلية لتقليل كل من خطأ التصنيف التجريبي وبُعد VC لدالة التصنيف. يتم وضع المعادلات ٦-٨ و٦-٢٣ معاً من خلال صياغة معادلة برمجية تربيعية:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \qquad (٦-٢٤)$$

بحيث تخضع للقيد:

$$y_i(w'x_i + b) \geq 1, \quad i = 1, \ldots, n.$$

٦-٣ التفسير الهندسي لصياغة الدعم الآلي المتجه (SVM) للمصنّف الخطي (Geometric Interpretation of the SVM Formulation for the Linear Classifier):

يوجد تفسير هندسي لـ $||w||$ في الدالة الهدف *(Objective function)* للمسألة البرمجية التربيعية في المعادلة ٦-٢٤ وهو أن $2/||w||$ تمثل المسافة للفضائين الجزئيين $w'x + b=1$ و$w'x + b=-1$ وتُسمى هذه المسافة هامش حد القرار أو هامش المصنّف الخطي، بحيث يكون المستقيم $w'x + b=0$ هو حد القرار. لإظهار هذا في الفضاء المتجه الثنائي الأبعاد لـ x دعونا نقوم بحساب مسافة المستقيمين المتوازيين $w'x + b=1$ و $w'x + b=-1$ في الشكل ٦-١. هذان المستقيمان المتوازيان اللذان يمكن تمثيلهما على النحو التالي:

$$w_1x_1 + w_2x_2 + b = 1 \qquad (٦\text{-}٢٥)$$

$$w_1x_1 + w_2x_2 + b = -1. \qquad (٦\text{-}٢٦)$$

المستقيم التالي:

$$w_2x_1 - w_1x_2 = 0 \qquad (٦\text{-}٢٧)$$

يمر عبر نقطة الأصل $(0,0)$، ويكون متعامداً على المستقيمات المعرفة في المعادلات ٦-٢٥ و٦-٢٦ لأن ميل المستقيمات المتوازية في المعادلات ٦-٢٥ و٦-٢٦ هو $-w_1/w_2$ وميل المستقيم في المعادلة ٦-٢٧ هو $-w_2 / w_1$ والذي هو المعكوس السالب لـ $-w_1 / w_2$. من خلال حل المعادلات ٦-٢٥ و٦-٢٧ لكل من x_1 و x_2 نحصل على إحداثيات نقطة البيانات حيث يتقاطع هذان المستقيمان:

$$\left(\frac{1-b}{w_1^2+w_2^2}w_1\,,\frac{1-b}{w_1^2+w_2^2}w_2\right)$$

من خلال حل المعادلات ٦-٢٦ و٦-٢٧ لكل من x_1 و x_2 نحصل على إحداثيات نقطة البيانات حيث يتقاطع هذان المستقيمان:

$$\left(\frac{-1-b}{w_1^2+w_2^2}w_1, \frac{-1-b}{w_1^2+w_2^2}w_2\right)$$

ثم نحسب المسافة بين نقطتي البيانات $\left(\frac{1-b}{w_1^2+w_2^2}w_1, \frac{1-b}{w_1^2+w_2^2}w_2\right)$ و $\left(\frac{-1-b}{w_1^2+w_2^2}w_1, \frac{-1-b}{w_1^2+w_2^2}w_2\right)$

$$d = \sqrt{\left(\frac{1-b}{w_1^2+w_2^2}w_1 - \frac{-1-b}{w_1^2+w_2^2}w_1\right)^2 + \left(\frac{1-b}{w_1^2+w_2^2}w_2 - \frac{-1-b}{w_1^2+w_2^2}w_2\right)^2}$$

$$= \frac{1}{w_1^2+w_2^2}\sqrt{2^2w_1^2+2^2w_2^2} = \frac{2}{\sqrt{w_1^2+w_2^2}} = \frac{2}{\|w\|}. \qquad (٦\text{-}٢٨)$$

ومن ثم، فإن تقليل قيمة $(1/2)\|w\|^2$ في دالة الهدف للمسألة البرمجية التربيعية في المعادلة ٦-٢٤ يكون بتعظيم هامش المصنْف الخطي أو أداء التعميم للمصنْف الخطي. يظهر الشكل ٦-١(a) و ٦-١(b) مصنْفين خطيين مختلفين بحدي قرار مختلفين يصنفان نقاط البيانات الثمان بشكل صحيح ولكن لهما هوامش مختلفة. يكون للمصنْف الخطي في الشكل ٦-١ (a) هامش أكبر، ومن المتوقع أن يكون له أداء تعميمي أفضل من ذلك التعميم في الشكل ٦-١ (b).

٦-٤ حل المسألة البرمجية التربيعية لمصنْف خطي

(Solution of the Quadratic Programming Problem for a Linear Classifier):

المسألة البرمجية التربيعية (*quadratic programming problem*) في الصيغة ٦-٢٤ لها دالة هدف تربيعية وقيد خطي بالنسبة لـ w وb، وتُسمى بمسألة التحسين المحدب (*Convex Optimization Problem*)، ويمكن حلها باستخدام طريقة مُضاعف لاقرينج (*Lagrange Multipliers*) للمسألة التالية:

$$\min_{w,b}\max_{\alpha\geq 0} L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(w'x_i + b) - 1\right] \quad (٦\text{-}٢٩)$$

بحيث تخضع للقيد:

$$\alpha_i[y_i(w'x_i + b) - 1] = 0 \quad i = 1, \dots, n \qquad (٦\text{-}٣٠)$$
$$\alpha_i \geq 0 \quad i = 1, \dots, n,$$

حيث α_i , $i=1,\dots,n$ ، هي مضاعفات لاقرينج غير السالبة، وتُعرَف المعادلتان المعرفتين في جزئية القيود بشرط كاروش-كوهن-توكر *(Karush – Kuhn – Tuker condition)* *(Burges,1998)* وتمثلان تحولاً لقيد المتراجحة في المعادلة ٦-٢٣. إن الحل للمعادلة ٦-٢٩ يكون عند النقطة الواصلة بين قمتين *(Saddle Point)* لـ $L= (w,b,\alpha)$، حيث يتم تصغير $L= (w,b,\alpha)$ بالنسبة لـ w وb وتعظيمها بالنسبة لـ α. يعطي تصغير $(1/2)||w||^2$ بالنسبة لــ w وb دالة الهدف في المعادلة ٦-٢٤. إن تصغير قيمة:

$$-\sum_{i=1}^{n} \alpha_i \left[y_i(w'x_i + b) - 1\right]$$

يكون بتعظيم قيمة:

$$\sum_{i=1}^{n} \alpha_i \left[y_i(w'x_i + b) - 1\right]$$

وذلك بالنسبة لـ α ويحقق $y_i(w'x_i + b) \geq 1$ - التي تمثل القيد في المعادلة ٦-٢٤، لأن $\alpha_i \geq 0$. عند النقطة حيث يتم تصغير $L(w,b,\alpha)$ بالنسبة لــ w وb، لدينا:

$$\frac{\partial L(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad or \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i \qquad (٣١-٦)$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (٣٢-٦)$$

لاحظ أنه يتم تحديد w فقط عن طريق نقاط البيانات التدريبية (x_i,y_i) ، والتي بها $\alpha_i < 0$. وتُسمى متجهات البيانات التدريبية والتي بها $\alpha_i > 0$ بالمتجهات الداعمة *(Support Vevtor)*. وباستخدام شرط كاروش-كوهن-توكر في المعادلة ٦-٣٠ وأي متجه دعم (x_i,y_i) بـ $\alpha_i < 0$، يكون لدينا:

$$y_i(w'x_i + b) - 1 = 0 \qquad (٣٣-٦)$$

من أجل تحقيق المعادلة ٦-٣٢. لدينا أيضاً:

$$y_i^2 = 1 \qquad (٣٤-٦)$$

لأن y_i تأخذ القيمة 1 أو 1-. نقوم بحل المعادلات ٦-٣٣ و٦-٣٤ لـ b ونحصل على:

$$b = y_i - w'x_i \qquad (٣٥-٦)$$

لأن:

$$y_i(w'x_i + b) - 1 = y_i(w'x_i + y_i - w'x_i) - 1 = y_i^2 - 1 = 0$$

ولحساب w باستخدام المعادلات ٦-٣١ و٦-٣٢ وحساب b باستخدام المعادلة ٦-٣٥، نحتاج أن نعرف قيم مضاعفات لاقرينج α. نقوم بتعويض المعادلات ٦-٣١ و٦-٣٢ داخل $L(w, b, \alpha)$ في الصيغة ٦-٢٩ للحصول على $L(\alpha)$:

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i' x_j - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i' x_j - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i$$
$$= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i' x_j \quad (٦\text{-}٣٦)$$

ومن ثم، فإن المسألة المزدوجة *(dual problem)* للمسألة البرمجية التربيعية في الصيغة ٦-٢٤ هي:

$$\max_{\alpha} L(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i' x_j \quad (٦\text{-}٣٧)$$

بشرط أن:

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$
$$\alpha_i[y_i(w' x_i + b) - 1] = 0 \;\; or \;\; \sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i' x_j + \alpha_i y_i b - \alpha_i = 0 \quad i = 1, \ldots, n$$
$$\alpha_i \geq 0 \quad i = 1, \ldots, n$$

وباختصار، فإنه يتم حل المصنّف الخطي للدعم الآلي المتجه SVM بالخطوات التالية:

١- حل مسألة التحسين في الصيغة ٦-٣٧ للحصول على α:

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j x_i' x_j$$

بشرط أن:

$$\sum_{i=1}^{n} \alpha_i \, y_i = 0$$

$$\sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j x_i' x_j + \alpha_i y_i b - \alpha_i = 0 \quad i = 1, \dots, n$$

$$\alpha_i \geq 0 \quad i = 1, \dots, n$$

٢- استخدم المعادلة ٦-٣١ للحصول على w:

$$w = \sum_{i=1}^{n} \alpha_i \, y_i x_i.$$

٣- استخدام المعادلة ٦-٣٥. ومتجه الدعم (x_i, y_i) للحصول على b:

$$b = y_i - w' x_i.$$

وتُعطى دالة قرار المصنّف الخطي بالمعادلة ٦-٢٢:

$$y = sign(w'x + b) = 1 \qquad \text{if } w'x + b \geq 1$$
$$y = sign(w'x + b) = -1 \qquad \text{if } w'x + b \leq -1.$$

أو بالمعادلة ٦-٤:

$$f_{w,b}(x) = sign(w'x + b) = sign\left(\sum_{i=1}^{n} \alpha_i\, y_i x_i' x + b\right).$$

لاحظ أن متجهات الدعم فقط والتي بها $\alpha_i > 0$ تسهم في حساب w، b ودالة قرار المصنًف الخطي.

المثال ٦-١:

حدّد المصنْف الخطي للدعم الآلي المتجه (SVM) لدالة AND في الجدول ٥-١، والتي يتم نسخها هنا في الجدول ٦-١ بحيث يكون $x=(x_1, x_2)$ هناك أربع نقاط من نقاط البيانات التدريبية في هذه المسألة. نقوم بصياغة وحل مسألة التحسين في الصيغة ٦-٢٤ على النحو التالي:

$$\min_{w_1,w_2,b} \frac{1}{2}\left[(w_1)^2 + (w_2)^2\right]$$

بشرط أن:

$$\begin{gathered} w_1 + w_2 - b \geq 1 \\ w_1 - w_2 - b \geq 1 \\ -w_1 + w_2 - b \geq 1 \\ w_1 + w_2 + b \geq 1. \end{gathered}$$

باستخدام شريط الأدوات المسمى (*Optimization*) في برنامج ماتلاب (*MATLAB®*)، نحصل على الحل الأمثل التالي لمسألة التحسين المذكورة آنفاً:

$$w_1=1,\ w_2=1,\ b=-1$$

وهذا يعني، أن لدينا:

$$w = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad b = -1.$$

هذا الحل يعطي دالة القرار في المعادلة ٦-٢٢ أو ٦-٤ كما يلي:

$$\begin{cases} y = sign\left(\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 1\right) = sign(x_1 + x_2 - 1) = 1 \quad \text{if}\, x_1 + x_2 - 1 \geq 1 \\ y = sign\left(\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 1\right) = sign(x_1 + x_2 - 1) = -1 \quad \text{if}\, x_1 + x_2 - 1 \leq -1 \end{cases}$$

أو

$$f_{w,b}(x) = sign(w'x + b) = sign\left(\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 1\right) = sign(x_1 + x_2 - 1).$$

الجدول (٦-١)

الدالة ***AND***

المخرجات Output	المدخلات Inputs		رقم سجل البيانات Data Point #
y	x_2	x_1	i
−1	−1	−1	1
−1	1	−1	2
−1	−1	1	3
1	1	1	4

يمكننا أيضاً صياغة مسألة التحسين في الصيغة ٦-٣٧:

$$\max_\alpha L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j x_i' x_j$$

$$= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}[\alpha_1\alpha_1 y_1 y_1 x_1' x_1 + \alpha_1\alpha_2 y_1 y_2 x_1' x_2$$

$$+\alpha_1\alpha_3 y_1 y_3 x_1' x_3 + \alpha_1\alpha_4 y_1 y_4 x_1' x_4 + \alpha_2\alpha_1 y_2 y_1 x_2' x_1 + \alpha_2\alpha_2 y_2 y_2 x_2' x_2$$

$$+\alpha_2\alpha_3 y_2 y_3 x_2' x_3 + \alpha_2\alpha_4 y_2 y_4 x_2' x_4 + \alpha_3\alpha_1 y_3 y_1 x_3' x_1 + \alpha_3\alpha_2 y_3 y_2 x_3' x_2$$

$$+\alpha_3\alpha_3 y_3 y_3 x_3' x_3 + \alpha_3\alpha_4 y_3 y_4 x_3' x_4 + \alpha_4\alpha_1 y_4 y_1 x_4' x_1 + \alpha_4\alpha_2 y_4 y_2 x_4' x_2$$
$$+\alpha_4\alpha_3 y_4 y_3 x_4' x_3 + \alpha_4\alpha_4 y_4 y_4 x_4' x_4]$$

$$= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\Big[\alpha_1\alpha_1(-1)(-1)[-1 \quad -1]\begin{bmatrix}-1\\-1\end{bmatrix}$$
$$+2\alpha_1\alpha_2(-1)(-1)[-1 \quad -1]\begin{bmatrix}-1\\1\end{bmatrix} + 2\alpha_1\alpha_3(-1)(-1)[-1 \quad -1]\begin{bmatrix}1\\-1\end{bmatrix}$$
$$+2\alpha_1\alpha_4(-1)(1)[-1 \quad -1]\begin{bmatrix}1\\1\end{bmatrix} + 2\alpha_2\alpha_2(-1)(-1)[-11]\begin{bmatrix}-1\\1\end{bmatrix}$$
$$+2\alpha_2\alpha_3(-1)(-1)[-11]\begin{bmatrix}1\\-1\end{bmatrix} + 2\alpha_2\alpha_4(-1)(1)[-11]\begin{bmatrix}1\\1\end{bmatrix}$$
$$+2\alpha_3\alpha_3(-1)(-1)[1 \quad -1]\begin{bmatrix}1\\-1\end{bmatrix} + 2\alpha_3\alpha_4(-1)(1)[1 \quad -1]\begin{bmatrix}1\\1\end{bmatrix}$$
$$+2\alpha_4\alpha_4(1)(1)[11]\begin{bmatrix}1\\1\end{bmatrix}\Big]$$

$$= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}(2\alpha_1^2 + 2\alpha_2^2 + 2\alpha_3^2 + 2\alpha_4^2 - 4\alpha_1\alpha_4 - 4\alpha_2\alpha_3)$$
$$= -\alpha_1^2 - \alpha_2^2 - \alpha_3^2 - \alpha_4^2 + 2\alpha_1\alpha_4 + 2\alpha_2\alpha_3 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$$
$$= -(\alpha_1 - \alpha_4)^2 - (\alpha_2 - \alpha_3)^2 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$$

بشرط أن:

$$\sum_{i=1}^{n} \alpha_i y_i = \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 + \alpha_4 y_4 = -\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 0$$

وتصبح $\left(\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j x_i' x_j + \alpha_i y_i b - \alpha_i = 0 \quad i = 1,2,3,4\right)$ كما يلي:

$$\alpha_1(-1)\Big[\alpha_1(-1)[-1 \quad -1]\begin{bmatrix}-1\\-1\end{bmatrix} + \alpha_2(-1)[-1 \quad 1]\begin{bmatrix}-1\\-1\end{bmatrix} + \alpha_3(-1)[1 \quad -1]\begin{bmatrix}-1\\-1\end{bmatrix}$$
$$\alpha_4(1)[1 \quad 1]\begin{bmatrix}-1\\-1\end{bmatrix}\Big] + \alpha_1(-1)b - \alpha_1 \quad or \quad -\alpha_1(-2\alpha_1 - 2\alpha_4) - \alpha_1 b - \alpha_1 = 0$$
$$\alpha_2(-1)\Big[\alpha_1(-1)[-1 \quad -1]\begin{bmatrix}-1\\1\end{bmatrix} + \alpha_2(-1)[-1 \quad 1]\begin{bmatrix}-1\\1\end{bmatrix} + \alpha_3(-1)[1 \quad -1]\begin{bmatrix}-1\\1\end{bmatrix}$$
$$\alpha_4(1)[1 \quad 1]\begin{bmatrix}-1\\1\end{bmatrix}\Big] + \alpha_2(-1)b - \alpha_2 \quad or \quad -\alpha_2(-2\alpha_2 - 2\alpha_3) - \alpha_2 b - \alpha_2 = 0$$

$$\alpha_3(-1)\left[\alpha_1(-1)[-1\quad -1]\begin{bmatrix}1\\-1\end{bmatrix}+\alpha_2(-1)[-1\quad 1]\begin{bmatrix}1\\-1\end{bmatrix}+\alpha_3(-1)[1\quad -1]\begin{bmatrix}1\\-1\end{bmatrix}\right.$$

$$\left.\alpha_4(1)[1\quad 1]\begin{bmatrix}1\\-1\end{bmatrix}\right]+\alpha_3(-1)b-\alpha_3\quad or\quad -\alpha_3(-2\alpha_2-2\alpha_3)-\alpha_3 b-\alpha_3=0$$

$$\alpha_4(1)\left[\alpha_1(-1)[-1\quad -1]\begin{bmatrix}1\\1\end{bmatrix}+\alpha_2(-1)[-1\quad 1]\begin{bmatrix}1\\1\end{bmatrix}+\alpha_3(-1)[1\quad -1]\begin{bmatrix}1\\1\end{bmatrix}\right.$$

$$\left.\alpha_4(1)[1\quad 1]\begin{bmatrix}1\\1\end{bmatrix}\right]+\alpha_4(1)b-\alpha_4\quad or\quad \alpha_4(2\alpha_1+2\alpha_4)+\alpha_4 b-\alpha_4=0$$

$$\alpha_i \geq 0\quad i=1,2,3,4$$

باستخدام شريط الأدوات المسمى (*Optimization*) في برنامج *MATLAB*® لحل مسألة التحسين المذكورة أعلاه، نحصل على الحل الأمثل:

$$\alpha_1=0,\quad \alpha_2=0.5,\quad \alpha_3=0.5,\quad \alpha_4=1,\quad b=-1,$$

وقيمة دالة الهدف تساوي 1.

تشير قيم مضاعفات لاقرينج إلى أن نقاط البيانات الثانية والثالثة والرابعة في الجدول ٦-١ هي متجهات الدعم. ثم نحصل بعد ذلك على w باستخدام المعادلة ٦-٣١:

$$w=\sum_{i=1}^{4}\alpha_i\, y_i x_i.$$

$$w_1=\alpha_1 y_1 x_{1,1}+\alpha_2 y_2 x_{2,1}+\alpha_3 y_3 x_{3,1}+\alpha_4 y_4 x_{4,1}$$
$$=(0)(-1)(-1)+(0.5)(-1)(-1)+(0.5)(-1)(1)+(1)(1)(1)=1$$

$$w_2=\alpha_1 y_1 x_{1,2}+\alpha_2 y_2 x_{2,2}+\alpha_3 y_3 x_{3,2}+\alpha_4 y_4 x_{4,2}$$
$$=(0)(-1)(-1)+(0.5)(-1)(1)+(0.5)(-1)(-1)+(1)(1)(1)=1$$

الحل الأمثل يتضمن بالفعل قيمة $b=-1$. نحصل على نفس قيمة b باستخدام المعادلة ٦-٣٥ ونقطة البيانات الرابعة كمتجه الدعم:

$$b = y_4 - w'x_4 = 1 - [1 \quad 1]\begin{bmatrix}1\\1\end{bmatrix} = -1.$$

يُعطي الحل الأمثل للمسألة المزدوجة للدعم الآلي المتجه *SVM* دالة القرار نفسه:

$$\begin{cases} y = sign\left([1 \quad 1]\begin{bmatrix}x_1\\x_2\end{bmatrix} - 1\right) = sign(x_1 + x_2 - 1) = 1 & \text{if } x_1 + x_2 - 1 \geq 1 \\ y = sign\left([1 \quad 1]\begin{bmatrix}x_1\\x_2\end{bmatrix} - 1\right) = sign(x_1 + x_2 - 1) = -1 & \text{if } x_1 + x_2 - 1 \leq -1 \end{cases}$$

أو

$$f_{w,b}(x) = sign(w'x + b) = sign\left([1 \quad 1]\begin{bmatrix}x_1\\x_2\end{bmatrix} - 1\right) = sign(x_1 + x_2 - 1).$$

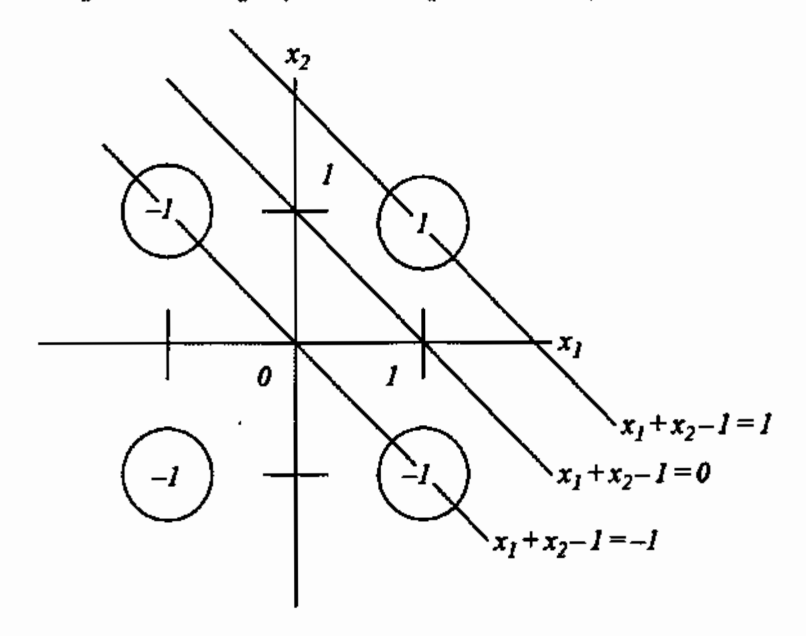
ومن ثم، فإن مسألة التحسين ومسألتها المزدوجة للدعم الآلي المتجه *SVM* لهذا المثال تعطى الحل الأمثل نفسه ودالة القرار. ويوضح الشكل ٦-٢ دالة القرار ومتجهات الدعم لهذه المسألة. دالة قرار الدعم الآلي المتجه *SVM* هي نفسها كما في شبكة الـ *ANN* لنفس المسألة الموضحة في الشكل ٥-١٠ في الفصل ٥.

العديد من الكتب وأوراق العمل في الدراسات العلمية تقدم موضوع الدعم الآلي المتجه *SVMs* باستخدام مسألة التحسين المزدوجة في الصيغة ٦-٣٧ ولكن من دون مجموعة القيود:

$$\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j x_i' x_j + \alpha_i y_i b - \alpha_i = 0 \quad i = 1, \ldots, n$$

الشكل (٦-٢)

دالة القرار ومتجهات الدعم للمصنّف الخطي الخاص بالدعم الآلي المتجه SVM في المثال ٦-١



كما يتضح من المثال ٦-١، من دون مجموعة القيود هذه، تصبح المسألة المزدوجة:

$$\max_{\alpha} -(\alpha_1-\alpha_4)^2-(\alpha_2-\alpha_3)^2+\alpha_1+\alpha_2+\alpha_3+\alpha_4$$

بشرط أن:

$$-\alpha_1-\alpha_2-\alpha_3+\alpha_4=0$$
$$\alpha_i \geq 0 \quad, i=1,2,3,4.$$

إذا وضعنا $\alpha_1=\alpha_4>0$، و$\alpha_2=\alpha_3=0$، التي تحقق جميع القيود، تصبح دالة الهدف بعد ذلك $max\ \alpha_1+\alpha_4$ التي تكون غير محدودة وغير مقيدة لأن كل من α_1 وα_4 يمكن أن تستمر في زيادة قيمها من دون حد. ومن ثم، فإنه ينبغي استخدام الصيغة ٦-٣٧ للمسألة المزدوجة مع المجموعة الكاملة من القيود.

٦-٥ صياغة الدعم الآلي المتجه (SVM) لمصنّف خطي ولمسألة قابلة للفصل بشكل غير خطي (SVM Formulation for a Linear Classifier and a Nonlinearly Separable Problem):

إذا تمّ تطبيق مصنّف خطي للدعم الآلي المتجه SVM على مسألة قابلة للفصل بشكل غير خطي (على سبيل المثال، دالة XOR المنطقية التي تمّ توضيحها في الفصل ٥)، فمن المتوقع أن لا يتم تصنيف كل نقطة بيانات في مجموعة بيانات العينة بشكل صحيح باستخدام المصنّف الخطي للدعم الآلي المتجه SVM. إن صيـــاغة دعم آلي متجـــه SVM لمصنّف خطي في الصيغة ٦-٢٤ يـمـكن أن يمتـد ليشمـل استخـــدام هامـش بسيـــط عن طريق إدخال مجموعـــة من المعـلمـات غيـر السالبـــة الإضـــافيـــة β_i $i=1,\dots,n$، في داخل صيغة الدعم الآلي المتجه SVM:

$$\min_{w,b,\beta} \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\beta_i\right)^k \qquad (٦\text{-}٣٨)$$

بشرط أن:

$$y_i(w'x_i + b) \geq 1 - \beta_i, \quad i = 1, \dots, n.$$
$$\beta_i \geq 0, \quad i = 1, \dots, n,$$

حيث إن $C > 0$ و $k \geq 1$: قيمتان محددتان سلفاً للحد من سوء تصنيف نقاط البيانات. إنَّ إدخال β_i في القيد في الصيغة رقم ٦-٣٨ يسمح بسوء تصنيف نقطة بيانات ما بمقدار β_i والتي تقيس مستوى الخطأ في التصنيف. إذا تم تصنيف نقطة بيانات ما بشكل صحيح، تصبح β_i صفراً. إن تقليل قيمة $C(\sum_{i=1}^{n}\beta_i)^k$ في دالة الهدف يكون بتقليل خطأ سوء التصنيف، في حين أن تقليل قيمة $(1/2)||w||^2$ في دالة الهدف يكون بتقليل بعد VC كما نوقش سابقاً.

باستخدام طريقة مضاعف لاقرينج، نقوم بتحويل الصيغة ٦-٣٨ إلى:

$$\min_{w,b,\beta}\max_{\alpha\geq 0,\gamma\geq 0} L(w,b,\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\beta_i\right)^k$$

$$-\sum_{i=1}^{n} \alpha_i \left[y_i(wx_i + b) - 1 + \beta_i\right] - \sum_{i=1}^{n} \gamma_i \beta_{i,} \qquad (٣٩\text{-}٦)$$

حيث γ_i ، $i=1,...,n$ ، هي مضاعفات لاقرينج غير السالبة. ويكون حل الصيغة ٦-٣٩ عند النقطة الواصلة بين قمتين لـ $L(w, b, \beta, \alpha, \gamma)$، حيث يتم تقليل $L(w, b, \beta, \alpha, \gamma)$، بالنسبة لـ w، b وβ، ويتم تعظيمها بالنسبة لـ α وγ. عند النقطة التي يتم فيها تقليل $L(w, b, \beta, \alpha, \gamma)$ بالنسبة لـ w، b وβ، يكون لدينا:

$$\frac{\partial L(w,b,\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \text{ or } w = \sum_{i=1}^{n} \alpha_i y_i x_i \qquad (٤٠\text{-}٦)$$

$$\frac{\partial L(w,b,\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (٤١\text{-}٦)$$

$$\frac{\partial L(w,b,\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} = \begin{cases} pC\left(\sum_{i=1}^{n} \beta_i\right)^{k-1} - \alpha_i - \gamma_i = 0 & i = 1,\dots,n \quad \text{if } k > 1 \\ C - \alpha_i - \gamma_i = 0 & i = 1,\dots,n \quad \text{if } k = 1 \end{cases} \qquad (٤٢\text{-}٦)$$

عندما تكون $k>1$، نرمز:

$$\delta = pC\left(\sum_{i=1}^{n} \beta_i\right)^{k-1} \quad \text{or} \quad \sum_{i=1}^{n} \beta_i = \left(\frac{\delta}{pC}\right)^{1/k-1}. \qquad (٤٣\text{-}٦)$$

يمكننا إعادة كتابة المعادلة ٦-٤٢ لتكون:

$$\begin{cases} \delta - \alpha_i - \gamma_i = 0 \quad \text{or} \quad \gamma_i = \delta - \alpha_i \quad i = 1, \dots, n \quad \text{if } k > 1 \\ C - \alpha_i - \gamma_i = 0 \quad \text{or} \quad \gamma_i = C - \alpha_i \quad i = 1, \dots, n \quad \text{if } k = 1 \end{cases} \quad (٦\text{-}٤٤)$$

شرط كاروش-كوهن-توكر للحل الأمثل للصيغة ٦-٣٩ يعطي:

$$\alpha_i[y_i(wx_i + b) - 1 + \beta_i] = 0. \quad (٦\text{-}٤٥)$$

باستخدام نقطة بيانات (x_i، y_i) والتي تُصنَّف بشكل صحيح بواسطة الدعم الآلي المتجه SVM، لدينا $\beta_i=0$ ولذلك يستند التالي إلى المعادلة ٦-٤٥:

$$b = y_i - w'x_i, \quad (٦\text{-}٤٦)$$

وهي المعادلة ٦-٣٥ نفسها. يتم استخدام المعادلتين ٦-٤٠ و٦-٤٦ لحساب w وb، على التوالي، إذا كانت α معروفة. نستخدم المسألة المزدوجة للصيغة ٦-٣٩ لتحديد α كما يلي.

عندما تكون $k = 1$ ، فإن التعويض بـ w وb، وγ في المعادلات ٦-٤٠، ٦-٤٤، و٦-٤٦، على التوالي، في الصيغة ٦-٣٩ يعطى:

$$\max_{\alpha \geq 0} L(\alpha) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\beta_i\right)^k - \sum_{i=1}^{n}\alpha_i[y_i(wx_i + b) - 1 + \beta_i] - \sum_{i=1}^{n}\gamma_i\beta_i$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\, y_i y_j x_i' x_j + C\sum_{i=1}^{n}\beta_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\sum_{j=1}^{n}\alpha_j y_j x_j' x_i + b\right) - 1 + \beta_i\right]$$

$$-\sum_{i=1}^{n}(C - \alpha_i)\beta_i = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\, y_i y_j x_i' x_j \quad (٦\text{-}٤٧)$$

بشرط أن:

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\alpha_i \leq C \quad i = 1, \dots, n$$
$$\alpha_i \geq 0 \quad i = 1, \dots, n.$$

القيد $\alpha_i \leq C$ يأتي من المعادلة ٦-٤٤:

$$C - \alpha_i - \gamma_i = 0 \quad \text{or} \quad C - \alpha_i = \gamma_i.$$

ولأن $\gamma_i \geq 0$ ، يكون لدينا $C \geq \alpha_i$.

عندما تكون $k > 1$، فإن التعويض بـ w وb، وγ في المعادلات ٦-٤٠، ٦-٤٤، و٦-٤٦، على التوالي، في الصيغة ٦-٣٩ يعطى:

$$\max_{\alpha \geq 0, \delta} L(\alpha) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \beta_i\right)^k - \sum_{i=1}^{n} \alpha_i [y_i(wx_i + b) - 1 + \beta_i] - \sum_{i=1}^{n} \gamma_i \beta_i$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, y_i y_j x_i' x_j + C\left(\sum_{i=1}^{n} \beta_i\right)^k - \sum_{i=1}^{n} \alpha_i \left[y_i\left(\sum_{j=1}^{n} \alpha_j y_j x_j' x_i + b\right) - 1 + \beta_i\right]$$

$$- \sum_{i=1}^{n} (\delta - \alpha_i)\beta_i = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, y_i y_j x_i' x_j - \frac{\delta^{\frac{p}{p-1}}}{(pC)^{\frac{1}{p-1}}}\left(1 - \frac{1}{p}\right) \qquad (٦\text{-}٤٨)$$

بشرط أن:

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\alpha_i \leq \delta \quad i = 1, \dots, n$$
$$\alpha_i \geq 0 \quad i = 1, \dots, n.$$

وتُعطَى دالة القرار للمصنِّف الخطي في المعادلة ٦-٢٢:

$$y = sign(w'x + b) = 1 \qquad \text{if}\, w'x + b \geq 1$$
$$y = sign(w'x + b) = -1 \quad \text{if}\, w'x + b \leq -1,$$

أو المعادلة ٦-٤:

$$f_{w,b}(x) = sign(w'x + b) = sign\left(\sum_{i=1}^{n} \alpha_i\, y_i x_i' x + b\right).$$

تسهم متجهات الدعم والتي فقط بها $\alpha_i>0$ في حساب قيم w، b، ودالة قرار المصنْف الخطي.

٦-٦ صياغة الدعم الآلي المتجه (SVM) لمصنْف غير خطي ومسألة قابلة للفصل بشكل غير خطي

(SVM Formulation for a Nonlinear Classifier and a Nonlinearly Separable Problem):

يتم توسيع الهامش البسيط للدعم الآلي المتجه SVM للمسألة القابلة للفصل بشكل غير خطى من خلال تحويل x ذات الأبعاد p في فضاء عدد أبعاده l حيث يمكن تصنيف x باستخدام المصنِّف الخطي. ويتم تمثيل عملية تحويل x كما يلي:

$$x \rightarrow \boldsymbol{\varphi}(x),$$

حيث إن:

$$\boldsymbol{\varphi}(x) = \big(h_1\varphi_1(x), \ldots, h_1\varphi_1(x)\big). \qquad (٦\text{-}٤٩)$$

وتصبح صياغة الهامش البسيط للدعم الآلي المتجه SVM:

عندما تكون $k = 1$،

$$\max_{\alpha\geq 0} L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j \boldsymbol{\varphi}(x_i)'\boldsymbol{\varphi}(x_j) \qquad (٦-٥٠)$$

بشرط أن:

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0$$
$$\alpha_i \leq C \quad i = 1, \ldots, n$$
$$\alpha_i \geq 0 \quad i = 1, \ldots, n$$

عندما تكون $k > 1$،

$$\max_{\alpha\geq 0,\delta} L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j \boldsymbol{\varphi}(x_i)'\boldsymbol{\varphi}(x_j) - \frac{\delta^{p/p-1}}{(pC)^{1/p-1}}\left(1-\frac{1}{p}\right) \qquad (٦-٥١)$$

بشرط أن:

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0$$
$$\alpha_i \leq \delta \quad i = 1, \ldots, n$$
$$\alpha_i \geq 0 \quad i = 1, \ldots, n$$

وبدالة قرار:

$$f_{w,b}(x) = sign\left(\sum_{i=1}^{n} \alpha_i\, y_i \boldsymbol{\varphi}(x_i)'\boldsymbol{\varphi}(x) + b\right). \qquad (٦-٥٢)$$

وإذا عرّفنا دالة كيرنل $K(x, y)$ على أنها:

$$K(x,y) = \boldsymbol{\varphi}(x)'\boldsymbol{\varphi}(y) = \sum_{i=1}^{l} h_i^2 \boldsymbol{\varphi}_i(x)'\boldsymbol{\varphi}_i(y), \qquad (٦\text{-}٥٣)$$

فإن صياغة الهامش البسيط للدعم الآلي المتجه SVM في المعادلات من ٦-٥٠ وحتى ٦-٥٢ تصبح:

عندما تكون $k = 1$،

$$\max_{\alpha \geq 0} L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, y_i y_j K(x_i, x_j) \qquad (٦\text{-}٥٤)$$

بشرط أن:

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0$$
$$\alpha_i \leq C \quad i = 1, \ldots, n$$
$$\alpha_i \geq 0 \quad i = 1, \ldots, n$$

عندما تكون $k > 1$،

$$\max_{\alpha \geq 0, \delta} L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, y_i y_j K(x_i, x_j) - \frac{\delta^{p/p-1}}{(pC)^{1/p-1}}\left(1 - \frac{1}{p}\right) \qquad (٦\text{-}٥٥)$$

بشرط أن:

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0$$
$$\alpha_i \leq \delta \quad i = 1, \ldots, n$$
$$\alpha_i \geq 0 \quad i = 1, \ldots, n$$

وبدالة القرار:

$$f_{w,b}(x) = sign\left(\sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b\right). \qquad (٦-٥٦)$$

يتطلب الهامش البسيط للدعم الآلي المتجه *SVM* في المعادلات ٦-٥٠ وحتى ٦-٥٢ تحويل $\varphi(x)$ ثم حل الدعم الآلي المتجه *SVM* في الفضاء المختار، في حين أن الهامش البسيط للدعم الآلي المتجه *SVM* في المعادلات من ٦-٥٤ وحتى ٦-٥٦ يستخدم دالة كيرنل $K(x, y)$ بشكل مباشر.

للعمل في الفضاء المختار باستخدام المعادلات ٦-٥٠ وحتى ٦-٥٢، يتم تقديم بعض الأمثلة على دوال التحويل لمتجه المدخلات x في فضاء ذي بعد واحد على النحو التالي:

$$\boldsymbol{\varphi}(x) = (1, x, \ldots, x^d) \qquad (٦-٥٧)$$

$$K(x,y) = \boldsymbol{\varphi}(x)'\boldsymbol{\varphi}(y) = 1 + xy + \cdots + (xy)^d.$$

$$\boldsymbol{\varphi}(x) = \left(\sin x, \frac{1}{\sqrt{2}}\sin(2x), \cdots, \frac{1}{\sqrt{i}}\sin(ix), \ldots\right) \qquad (٦-٥٨)$$

$$K(x,y) = \boldsymbol{\varphi}(x)'\boldsymbol{\varphi}(y) = \sum_{i=1}^{\infty}\frac{1}{i}\sin(ix)\sin(iy) = \frac{1}{2}\log\left|\frac{\sin(x+y/2)}{\sin(x-y/2)}\right|$$

$$x, y \in [0, \pi].$$

وفيما يلي يتم إعطاء مثال على دالة تحويل لمتجه مدخلات $x = (x_1, x_2)$ في فضاء ذي بعدين:

$$\boldsymbol{\varphi}(x) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2\right) \qquad (٦-٥٩)$$

$$K(x,y) = \boldsymbol{\varphi}(x)'\boldsymbol{\varphi}(y) = (1 + xy)^2.$$

وفيما يلي يتم إعطاء مثال على دالة تحويل لمتجه المدخلات $x = (x_1, x_2, x_3)$ في فضاء ثلاثي الأبعاد:

$$\boldsymbol{\varphi}(x) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2,\ x_2^2,\ x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3\right) \qquad (٦٠-٦)$$

$$K(x,y) = \boldsymbol{\varphi}(x)'\boldsymbol{\varphi}(y) = (1 + xy)^2.$$

يمكن استخدام تحليل المكون الرئيسي (*principle component analysis*) الوارد في الفصل ١٤ لاستخراج المكونات الرئيسية لبناء $\varphi(x)$. لكن، قد لا تعطي المكونات الرئيسة بالضرورة الخواص أو الصفات المناسبة التي تؤدي إلى مصنِّف خطي في الفضاء المختار.

بالنسبة لدوال التحول في المعادلات من ٦-٥٧ وحتى ٦-٦٠، من الأسهل حساب دالة كيرنل مباشرةً بدلاً من البدء بحساب دوال التحويل والعمل في الفضاء المختار لأن الدعم الآلي المتجه SVM يمكن حله باستخدام دالة كيرنل مباشرةً. وفيما يلي ترد بعض الأمثلة لدوال كيرنل:

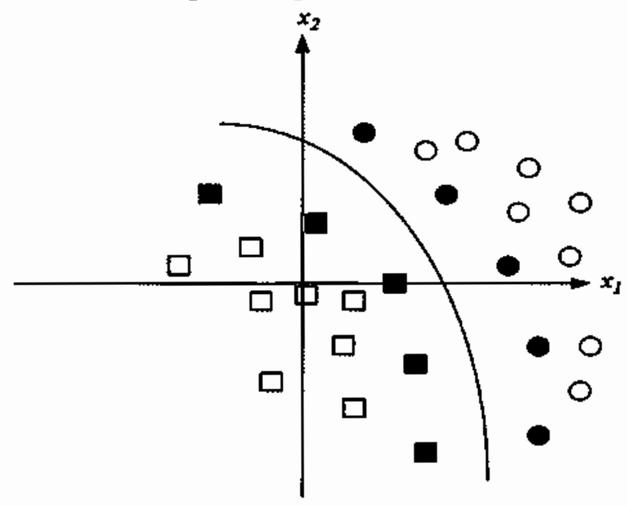
$$K(x,y) = (1 + xy)^2 \qquad (٦١-٦)$$

$$K(x,y) = e^{\frac{\|x-y\|^2}{2\sigma^2}} \qquad (٦٢-٦)$$

$$K(x,y) = \tanh(\rho xy - \theta). \qquad (٦٣-٦)$$

تعطي دوال كيرنل في المعادلات من ٦-٦١ وحتى ٦-٦٣ دالة قرار كثيرة الحدود (*polynomial decision function*) كما هو مبين في الشكل ٦-٣، ودالة القاعدة الدائرية لقوسشيان (*Gaussian Radial Basis function*) كما هو مبين في الشكل ٦-٤، والشبكة العصبية الصناعية ذات التغذية الأمامية الأحادية الطبقة (*perception*) متعددة السنوات لبعض قيم ρ و θ.

الشكل (٦-٣)
دالة قرار كثيرة الحدود في فضاء ثنائي الأبعاد

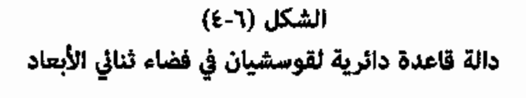


غالباً ما يتم استخدام عملية الجمع (*addition*) وعملية الضرب الممتد (*tensor Product*) لدوال كيرنل لبناء دوال كيرنل أكثر تعقيداً على النحو التالي:

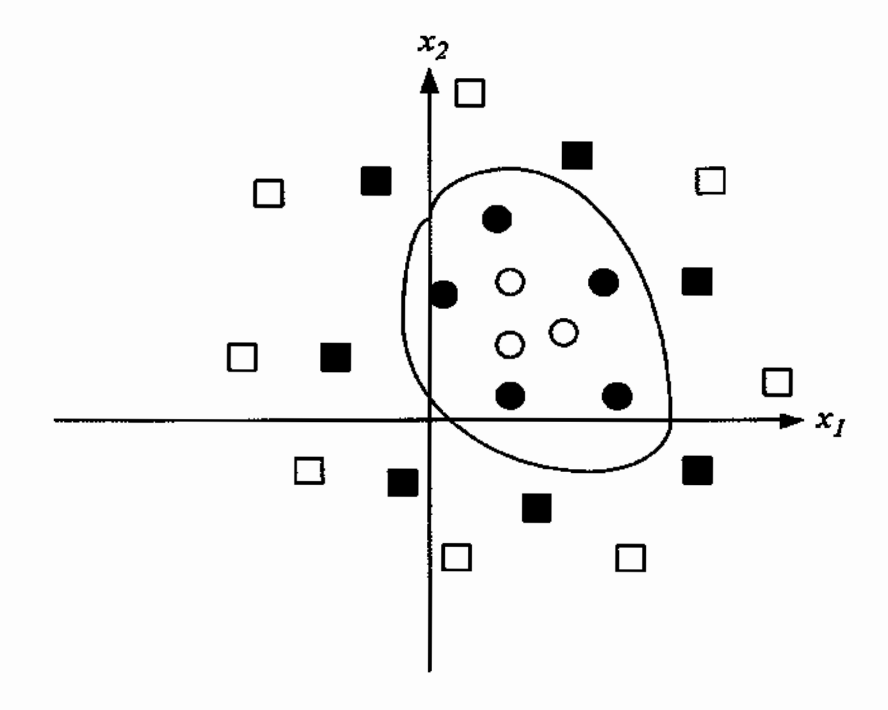
$$K(x,y) = \sum_i K_i(x,y) \qquad (٦٤-٦)$$

$$K(x,y) = \prod_i K_i(x,y). \qquad (٦٥-٦)$$

الشكل (٦-٤)



دالة قاعدة دائرية لقوسشيان في فضاء ثنائي الأبعاد



٦-٧ طرق استخدام الدعم الآلي المتجه (SVM) لمسائل التصنيف متعددة الفئات (Methods of Using SVM for Multi-Class Classification Problems):

الدعم الآلي المتجه SVM الموضح في الأجزاء السابقة هو لمصنِّف ثنائي يتعامل مع فئتين مستهدفتين. بالنسبة إلى مسألة تصنيف بأكثر من فئتين مستهدفتين، هناك العديد من الأساليب التي يمكن استخدامها لبناء مصنِّف ثنائي أولاً ثم الجمع بين المصنِّفات الثنائية للتعامل مع فئات مستهدفة متعددة. لنفترض أن الفئات المستهدفة هي $T_1, T_2, \dots, T_s$. في الأسلوب واحد مقابل واحد ($One - Versus - One$)، يتم بناء مصنِّف ثنائي لكل زوج من الفئات المستهدفة، T_i مقابل T_j، بحيث $i \neq j$. من بين الفئات المستهدفة التي تنتجها جميع المصنِّفات الثنائية لمتجه مدخلات معين، فإنه يتم أخذ الفئة المستهدفة المسيطرة كفئة مستهدفة نهائية لمتجه المدخلات. في الأسلوب واحد مقابل الكل ($One - Versus - all$)،

لنفترض أن مصنِّفاً ثنائياً يتم بناؤه ليتم تمييز كل فئة مستهدفة T_i من جميع الفئات المستهدفة الأخرى التي يتم اعتبارها معاً فئة مستهدفة أخرى ($NOT- T_i$). إذا كانت جميع المصنِّفات الثنائية ينتج عنها حصيلة تصنيف متسقة لمتجه مدخلات معين من ضمنها مصنِّف ثنائي واحد يعطي T_i وجميع المصنِّفات الأخرى تعطي فئات مستهدفة ليست T_j بحيث أن $i \neq j$، فإن الفئة المستهدفة النهائية لمتجه المدخلات تكون T_i. لكن إذا كانت جميع المصنِّفات الثنائية ينتج عنها حصيلة تصنيف غير متسقة لمتجه مدخلات معين، فإنه من الصعب تحديد الفئة المستهدفة النهائية لمتجه المدخلات. على سبيل المثال، قد يكون هناك فئتان مستهدفتان T_i، وT_j؛ بحيث $j \neq i$ في حصيلة التصنيف، وأنه من الصعب تحديد ما إذا كانت الفئة المستهدفة النهائية هي T_i، أو T_j. فإن أسلوب ترميز مخرجات تصحيح الخطأ (*Error – Correction Output Coding Method*) يولِّد رمزاً ثنائياً فريداً يتألف من خوينتين أو بت ثنائي (*binary bits*) لكل فئة مستهدفة، ثم تبني مصنِّفاً ثنائياً لكل خوينة أو بت ثنائي واحد، ثم تأخذ الفئة المستهدفة ذات السلسة من البتات الثنائية الأقرب إلى السلسلة الناتجة من البتات الثنائية من جميع المصنِّفات الثنائية. على الرغم من ذلك، لا يوجد طريقة مباشرة واضحة لتوليد رمز ثنائي فريد لكل فئة مستهدفة بحيث تؤدي مجموعة الرموز الثنائية الناتجة لجميع الفئات المستهدفة إلى الحد الأدنى من الخطأ في التصنيف لسجلات البيانات التدريبية أو الاستكشافية.

٦-٨ مقارنة بين الشبكة العصبية الصناعية (ANN) والدعم الآلي المتجه (SVM) (Comparison of ANN and SVM):

علمنا أن تعلُّم الشبكة العصبية الصناعية *ANN* كما هو موضح في الفصل ٥، يتطلب البحث عن الأوزان والتحيزات لشبكة الـ*ANN* نحو الحد الأدنى من خطأ تصنيف نقاط البيانات التدريبية، على الرغم من أن عملية البحث قد تنتهي بقاع محلي (*local minimum*). يتم حل الدعم الآلي المتجه *SVM* للحصول على الحل الأمثل على مستوى شامل. ولكن، بالنسبة للمصنِّف غير الخطي والمسألة القابلة للفصل بشكل غير خطي، غالباً ما يكون غير مؤكد ما هي دالة كيرنل الأصح لتحويل المسألة غير الخطية إلى مسألة قابلة للفصل خطياً لأن دالة التصنيف المناسبة غير معروفة. دون وجود دالة كيرنل مناسبة، فقد ينتهي بنا الأمر إلى استخدام دالة كيرنل غير مناسبة، ومن ثم الوصول إلى حل بخطأ تصنيفي أكبر من ذلك الناتج عن الحل الأمثل الشامل عند استخدام دالة كيرنل مناسبة. ومن ثم،

فاستخدام الدعم الآلي المتجه *SVM* لمصنِّف غير خطي ولمسألة قابلة للفصل بشكل غير خطي يستلزم البحث عن دالة كيرنل جيدة لتصنيف البيانات التدريبية من خلال التجربة والخطأ، تماماً كما أن تعلم شبكة عصبية صناعية *ANN* يستلزم تحديد تركيب مناسب لشبكة الـ*ANN* (أي عدد الوحدات الخفية) من خلال التجربة والخطأ. وعلاوةً على ذلك، فإن حساب:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\, y_i y_j x_i' x_j$$

أو

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\, y_i y_j K(x_i,x_j)$$

في دالة الهدف للدعم الآلي المتجه *SVM* لمجموعة كبيرة من البيانات التدريبية (على سبيل المثال، مجموعة تحتوي على ٥٠,٠٠٠ سجل بيانات تدريب) يتطلب حساب 2.5×10^9 حد ومساحة ذاكرة كبيرة، ومن ثم يؤدي إلى تكلفة حاسوبية (***computational cost***) كبيرة. يطبق أوسونا وآخرون (***Osuna et al., 1997***) الدعم الآلي المتجه *SVM* لمسألة كشف الوجه (***Face Detection Problem***)، ويبين أن أداء تصنيف الدعم الآلي المتجه *SVM* يظهر قريباً من أداء التصنيف باستخدام شبكة الـ *ANN* والمطور من قبل كل من سونغ وبوجيو (*Sung and Poggio, 1998*).

٦-٩ البرمجيات والتطبيقات (Software and Applications):

يدعم برنامج *MATLAB®* (*www.mathworks.com*) الدعم الآلي المتجه *SVM*. يمكن استخدام شريط الأدوات المسمى (*Optimization*) في برنامج *MATLAB®* لحل أي مشكلة تحسين باستخدام الدعم الآلي المتجه *SVM*. قام أوسونا وآخرون (***Osuna et al., 1997***) بتطبيق الدعم الآلي المتجه *SVM* لمسألة كشف الوجه. هناك العديد من

التطبيقات الأخرى للدعم الآلي المتجه *SVM* مذكورة في المراجع العلمية (*www.support-vector-machines.org*).

التمارين (Exercises):

٦-١ قم بتحديد المصنِّف الخطي للدعم الآلي المتجه *SVM* للدالة *OR* في الجدول ٥-٢ باستخدام صياغة الدعم الآلي المتجه *SVM* لمصنِّف خطي في الصيغتين ٦-٢٤ و٦-٢٩.

٦-٢ قم بتحديد المصنِّف الخطي للدعم الآلي المتجه *SVM* للدالة *NOT* باستخدام صياغة الدعم الآلي المتجه *SVM* لمصنِّف خطي في الصيغتين ٦-٢٤ و٦-٢٩. وترد مجموعة البيانات التدريبية لدالة *NOT*، *y= NOT*، فيما يلي:

مجموعة البيانات التدريبية:

Y	X
1	-1
-1	1

٦-٣ قم بتحديد المصنِّف الخطي للدعم الآلي المتجه *SVM* لدالة تصنيف مع البيانات التدريبية التالية، وذلك باستخدام صياغة الدعم الآلي المتجه *SVM* لمصنِّف خطي في الصيغتين ٦-٢٤ و٦-٢٩.

مجموعة البيانات التدريبية:

y	x_3	x_2	x_1
0	-1	-1	-1
0	1	-1	-1
0	-1	1	-1
1	1	1	-1
0	-1	-1	1
1	1	-1	1
1	-1	1	1
1	1	1	1

٧- مصنِّف أقرب k- مجاور والتعنقُد المراقب

k-Nearest Neighbor Classifier and Supervised Clustering

يستعرض هذا الفصل طريقتين للتصنيف، وهما: مصنِّف أقرب *k*- مجاور (*k-nearst neighbor classifier*)، والتعنقُد المراقب (*supervised clustering*)، والذي يتضمن مصنِّف أقرب *k*- مجاور كجزء من خوارزميته. كما يستعرض هذا الفصل بعض التطبيقات المتعلقة بالتعنقُد المراقب مع المراجع الخاصة به.

٧-١ مصنِّف أقرب *k*-مجاور (***k*-Nearest Neighbor Classifier**):

بالنسبة لنقطة أو سجل بيانات x_i بعدد p من متغيرات الخاصية (*attribute variables*):

$$x_i = \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{bmatrix}$$

ومتغير هدف (*target variable*) واحد، y، الذي يحتاج إلى أن يتم تحديد قيمتة النوعية، فإن مصنِّف أقرب *k*- مجاور يحدد أولاً موقع عدد *k* من نقاط أو سجلات البيانات الأكثر تشابهاً لـ (أوالأقرب إلى) نقطة البيانات هذه، كأقرب *k*- مجاور لنقطة البيانات، ثم يقوم المصنفُ باستخدام الفئات الهدف (*target classes*) للمجاورين الأقرب والتي عددها *k* لتحديد الفئة الهدف لنقطة البيانات. لتحديد أقرب *k*- مجاور لنقطة البيانات، نحتاج إلى استخدام مقياس للتشابه أو الاختلاف بين نقاط البيانات. يوجد العديد من مقاييس التشابه أو الاختلاف، بما في ذلك المسافة الإقليدية (*Euclidean distance*)، ومسافة مينكوسكي (*Minkowski distance*)، ومسافة هامينغ (*Hamming distance*)، ومعامل ارتباط بيرسون (*Pearson's correlation coefficient*)، وتشابه جيب التمام (جتا) (*Cosine similarity*)، والتي سيتم شرحها في هذا الجزء.

يتم تعريف المسافة الإقليدية على أنها:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p}(x_{i,l} - x_{j,l})^2} \quad, i \neq j. \qquad (١-٧)$$

المسافة الإقليدية هي مقياس الاختلاف بين نقطتي بيانات x_i وx_j. كلما كانت قيمة المسافة الإقليدية أكبر، كان الاختلاف بين نقطتي البيانات أكبر، ومن ثم متباعدتان إحداهما عن الأخرى بشكل أكبر ومنفصلتان أكثر في فضاء بيانات عدد أبعاده p.

يتم تعريف مسافة مينكوسكي (*Minkowski distance*) على أنها:

$$d(x_i, x_j) = \left(\sum_{l=1}^{p}|x_{i,l} - x_{j,l}|^r\right)^{1/r} \quad, i \neq j. \qquad (٢-٧)$$

مسافة مينكوسكي هي أيضاً مقياس للاختلاف. إذا وضعنا $r = 2$، فإن قيمة مسافة مينكوسكي تعطي نفس قيمة المسافة الإقليدية. إذا وضعنا $r = 1$، و يأخذ كل متغير من متغيرات الخاصية قيمة ثنائية، فإن قيمة مسافة مينكوسكي تعطي نفس قيمة مسافة هامينغ التي تقوم بتعداد عدد الخوينات أو البتات (*bits*) المختلفة بين سلسلتين ثنائيتين (*two binary strings*).

عندما يتم استخدام مقياس مسافة مينكوسكي، قد يكون لمتغيرات الخاصية المختلفة متوسطات (*means*)، وتباينات (*variances*) ونطاقات (*ranges*) مختلفة، وتجلب مستويات مختلفة في عملية حساب المسافة. على سبيل المثال، القيم الخاصة بمتغير من متغيرات الخاصية، x_i قد تتراوح من 0 إلى 10، في حين أن قيم متغير خاصية آخر، x_j قد تتراوح من 0-1. قيمتان للمتغير x_i ولتكن 1 و8، تعطي الفرق المطلق 7، في حين أن القيمتين لـ x_j 0.1 و0.8، تعطي الفرق المطلق 0.7. عندما تُستخدم كلٌ من القيمتين 7 و0.7 في جمع الفروقات بين نقطتي بيانات على مستوى جميع متغيرات الخاصية في المعادلة ٧-٢،

يكون الفرق المطلق على مستوى x_j غير ذي صلة عند مقارنته بالفرق المطلق على مستوى x_i ومن ثم، قد يكون من الضروري القيام بالتطبيع (*normalization*) قبل استخدام مقياس مسافة مينكوسكي. ويمكن استخدام عدة أساليب للتطبيع. وتستخدم إحدى أساليب التطبيع الصيغة التالية لتطبيع المتغير x والحصول على المتغير المُطبع z بمتوسط قيمته صفر، وتباين قيمته 1:

$$z = \frac{x - \bar{x}}{s}, \qquad (٣\text{-}٧)$$

حيث $\bar{x}$ وs هما متوسط العينة والانحراف المعياري للعينة الخاصة بالمتغير x على التوالي.

طريقةٌ أخرى للتطبيع تَستخدم الصيغة التالية لتطبيع المتغير x وإنتاج المتغير المطبع z مع القيم التي تتراوح من $[0,1]$:

$$z = \frac{x_{max} - x}{x_{max} - x_{min}}. \qquad (٤\text{-}٧)$$

يتم تنفيذ التطبيع من خلال تطبيق نفس طريقة التطبيع لجميع متغيرات الخاصية. وتُستخدم متغيرات الخاصية المطبعة لحساب مسافة مينكوسكي.

يُعرَف ما يلي معامل الارتباط بيرسون ρ:

$$\rho_{x_i x_j} = \frac{S_{x_i x_j}}{S_{x_i} S_{x_j}}, \qquad (٥\text{-}٧)$$

حيث s_{xixj}، وs_{xi} و s_{xj} تمثل التغاير (*covariance*) المُقدَّر لـ x_i وx_j حيث الانحراف المعياري المقدر لـ x_i والانحراف المعياري المقدرلـ x_j على التوالي، ويتم حسابها باستخدام عينة من نقاط البيانات n كما يلي:

$$S_{x_i x_j} = \frac{1}{n-1} \sum_{l=1}^{p} \left(x_{i,l} - \bar{x_i}\right)\left(x_{j,l} - \bar{x_j}\right) \qquad (\text{٧-٦})$$

$$S_{x_i} = \sqrt{\frac{1}{n-1} \sum_{l=1}^{p} \left(x_{i,l} - \bar{x_i}\right)^2} \qquad (\text{٧-٧})$$

$$S_{x_j} = \sqrt{\frac{1}{n-1} \sum_{l=1}^{p} \left(x_{j,l} - \bar{x_j}\right)^2} \qquad (\text{٧-٨})$$

$$\bar{x_i} = \frac{1}{n} \sum_{l=1}^{p} x_{i,l} \qquad (\text{٧-٩})$$

$$\bar{x_j} = \frac{1}{n} \sum_{l=1}^{p} x_{j,l}. \qquad (\text{٧-١٠})$$

يقع معامل ارتباط بيرسون في النطاق *[-1,1]* وهو مقياس للتشابه بين نقطتي البيانات x_i وx_j. كلما كبرت قيمة معامل الارتباط بيرسون، زاد الارتباط أو التماثل بين أو التشابه بين نقطتي البيانات ويرد وصف أكثر تفصيلاً لمعامل ارتباط بيرسون في الفصل ١٤.

ويَعُد مقياس تشابه جيب التمام (جتا) نقطتي البيانات x_i وx_j على أنهما متجهان في فضاء عدد أبعاده p ويستخدم جيب تمام الزاوية θ بين المتجهين لقياس التشابه بين نقطتي البيانات على النحو التالي:

$$\cos(\theta) = \frac{x_i' x_j}{\|x_i\| \|x_j\|}, \qquad (\text{٧-١١})$$

حيث $\|x_i\|$ و$\|x_j\|$ تمثل طولي المتجهين، ويتم حسابها على النحو التالي:

$$\|x_i\| = \sqrt{x_{i,1}^2 + \cdots + x_{i,p}^2} \quad (١٢-٧)$$

$$\|x_j\| = \sqrt{x_{j,1}^2 + \cdots + x_{j,p}^2} \quad (١٣-٧)$$

عندما $\theta = 0^o$ ، فهذا يعني، أن المتجهين الاثنين يشيران إلى نفس الاتجاه، $cos(\theta)=1$. عندما $\theta=180^o$، فهذا يعني، أن المتجهين الاثنين يشيران إلى اتجاهين متعاكسين، $cos(\theta)=-1$. عندما $\theta=90^o$، أو 270^o، فهذا يعني، أن المتجهين الاثنين متعامدين، $cos(\theta)=0$. ومن ثم، مثل معامل ارتباط بيرسون، فإن مقياس تشابه جيب التمام (جتا) يعطي قيمة في النطاق [-1,1]، وهو مقياس التشابه بين نقطتي البيانات x_i وx_j كلما كانت قيمة مقياس تشابه جيب التمام (جتا) أكبر، كانت نقطتا البيانات متشابهتين. ويرد وصف أكثر تفصيلاً لحساب الزاوية بين متجهي بيانات في الفصل ١٤.

لتصنيف نقطة بيانات x يتم حساب مقدار تشابه نقطة البيانات x لكل من نقاط البيانات n في مجموعة البيانات التدريبية باستخدام مقياس محدد للتشابه أو الاختلاف. من بين نقاط البيانات n في مجموعة البيانات التدريبية، فإن نقاط البيانات k والتي تكون أكثر تشابهاً لنقطة البيانات x يتم اعتبارها أقرب k- مجاور لـ x وتُؤخذ فئة الهدف المهيمنة والخاصة بأقرب k- مجاور كفئة الهدف لـ x وبعبارة أخرى، فإن مصنف أقرب k- مجاور يستخدم قاعدة تصويت الأغلبية لتحديد الفئة الهدف لـ x. على سبيل المثال، افترض أنه لتصنيف نقطة البيانات x يكون لدينا ما يلي:

- يتم وضع k عند 3.
- يأخذ المتغير الهدف واحد من فئتي الهدف: A وB.
- يكون لاثنين من أقرب ٣- مجاور الفئة الهدف A.

يقوم مصنف أقرب ٣- مجاور بإسناد القيمة A كفئة هدف لنقطة البيانات x.

المثال (٧-١):

استخدام مصنِّف أقرب ٣- مجاور، ومقياس المسافة الإقليدية للاختلاف لتصنيف ما إذا كان نظام التصنيع متعطل باستخدام قيم متغيرات الجودة التسعة. حيث تعطي مجموعة البيانات التدريبية في الجدول ٧- ١ جزءاً من مجموعة البيانات في الجدول ١-٤، وتتضمن تسع حالات أعطال منفردة، وحالة واحدة بدون أعطال في نظام التصنيع. بالنسبة لسجل البيانات رقم (i)، هناك تسعة متغيرات من متغيرات الخاصية لجودة وحدات المنتج، $(x_{i1},$ $\dots,x_{i9})$، ومتغير هدف واحد y_i لعطل النظام. يعطي الجدول ٧- ٢ حالات الاختبار لبعض الحالات متعددة الأعطال.

لنقطة البيانات الأولى في مجموعة البيانات الاختيارية x=(1,1,0,1,1,0,1,1,1) وكانت المسافات الإقليدية لنقطة البيانات هذه وصولاً إلى نقاط البيانات العشرة في مجموعة البيانات التدريبية هي: 1.73، 2، 2.45، 2.24، 2، 2.65، 2.45، 2.45، 2.45، 2.65، على التوالي. على سبيل المثال، المسافة الأقليدية بين x ونقطة البيانات الأولى في مجموعة البيانات التدريبية x=(1,0,0,0,1,0,1,0,1) هي:

$$d(x_1,x)=\sqrt{\begin{matrix}(1-1)^2+(0-1)^2+(0-0)^2+(0-1)^2+(1-1)^2+(0-0)^2\\+(1-1)^2+(0-1)^2+(1-1)^2\end{matrix}}=\sqrt{3}=1.73$$

أقرب ٣-مجاورات لـ x هي x_1 x_2 وx_5 في مجموعة البيانات التدريبية التي تأخذ جميعها الفئة الهدف 1 مما يعني نظاماً معطلاً. ومن ثم، يتم إسناد الفئة الهدف 1 لنقطة البيانات الأولى في مجموعة البيانات الاختيارية. حيث إنه في مجموعة البيانات الاختبارية، هناك نقطة بيانات واحدة فقط بالفئة الهدف صفر، فإن أقرب ٣-مجاور لكل نقطة بيانات في مجموعة البيانات الاختيارية، يكون لها على الأقل نقطتا البيانات التي فئتها الهدف 1 ، مما ينتج عنه قيمه للفئة الهدف تساوى 1 لكل نقطة البيانات في مجموعة البيانات الاختيارية. إذا حاولنا تصنيف نقطة البيانات رقم 10 بفئة هدف حقيقية تساوى صفراً في مجموعة البيانات التدريبية، فإن أقرب ٣-مجاور لهذه النقطة هي نقطة البيانات نفسها، بالإضافة لنقطتي بيانات أخريين فئتها الهدف تساوى1، مما يجعل فئة الهدف تساوى 1 لنقطة البيانات رقم 10 في مجموعة البيانات التدريبية، والذي يختلف عن الفئة الهدف الحقيقية لنقطة البيانات هذه.

الجدول (٧-١)

مجموعة البيانات التدريبية الخاصة بالكشف عن الأعطال بنظام التصنيع

متغير الهدف Target Variable	Attribute Variables - متغيرات الخاصية									
عطل النظام (System Fault), y_i	Quality of Parts - جودة وحدات المنتج									رقم الحالة i Instance i (الآلة المعطلة) (Faulty Machine
	x_{i9}	x_{i8}	x_{i7}	x_{i6}	x_{i5}	x_{i4}	x_{i3}	x_{i2}	x_{i1}	
1	1	0	1	0	1	0	0	0	1	1 (M1)
1	0	1	0	0	0	1	0	1	0	2(M2)
1	0	1	1	1	0	1	1	0	0	3(M3)
1	0	1	0	0	0	1	0	0	0	4(M4)
1	1	0	1	0	1	0	0	0	0	5(M5)
1	0	0	1	1	0	0	0	0	0	6(M6)
1	0	0	1	0	0	0	0	0	0	7(M7)
1	0	1	0	0	0	0	0	0	0	8(M8)
1	1	0	0	0	0	0	0	0	0	9(M9)
0	0	0	0	0	0	0	0	0	0	10(none)

الجدول (٧-٢)

مجموعة البيانات الاختيارية الخاصة بالكشف عن الأعطال بنظام التصنيع ونتائج التصنيف في الأمثلة ٧-١ و٧-٢

متغير الهدف- Target Variable (أعطال النظام) (System Faults y_i) القيمة المصنفة (Classified Value)	القيمة الفعلية (True Value)	متغيرات الخاصية - Attribute Variables (جودة وحدات المنتج - Quality of Parts) x_{i9}	x_{i8}	x_{i7}	x_{i6}	x_{i5}	x_{i4}	x_{i3}	x_{i2}	x_{i1}	رقم الحالة i Instance i (الآلة المعطلة Faulty Machine)
1	1	1	1	1	0	1	1	0	1	1	1 (M1,M2)
1	1	0	1	1	1	0	1	1	1	0	2(M2,M3)
1	1	1	0	1	1	1	0	1	0	1	3(M1,M3)
1	1	1	1	1	0	1	1	0	0	1	4(M1,M4)
1	1	1	0	1	1	1	0	0	0	1	5(M1,M6)
1	1	0	1	1	1	0	1	0	1	0	6(M2,M6)
1	1	0	1	1	0	1	1	0	1	0	7(M2,M5)
1	1	1	0	1	1	1	0	1	0	0	8(M3,M5)
1	1	0	1	1	0	0	1	0	0	0	9(M4,M7)
1	1	0	1	1	0	1	0	0	0	0	10(M5,M8)
1	1	1	1	1	1	0	1	1	0	0	11(M3,M9)
1	1	1	1	1	0	1	0	0	0	1	12(M1,M8)
1	1	1	1	1	1	1	1	1	1	1	13(M1,M2,M3)
1	1	1	1	1	1	1	1	1	1	0	14(M2,M3,M5)
1	1	1	1	1	1	0	1	1	1	0	15(M2,M3,M9)
1	1	1	1	1	1	1	0	0	0	1	16(M1,M6,M8)

ولكن، إذا وضعنا $k=1$ لهذا المثال، فإن مصنِّف أقرب ١- مجاور يُسند فئة الهدف الصحيحة لكل نقطة بيانات في مجموعة البيانات التدريبية لأن كل نقطة بيانات في مجموعة البيانات التدريبية لها نفسها كأقرب ١- مجاور ويسند أقرب ١- مجاور أيضاً فئة الهدف الصحيحة ١ لكل نقطة البيانات في مجموعة البيانات الاختيارية لأن نقطة البيانات رقم 10 في مجموعة البيانات التدريبية هي نقطة البيانات الوحيدة ذات الفئة الهدف صفر،

والمتغيرات الخاصية الخاصة بها تحتوي على قيم الصفر، مما يجعل نقطة البيانات رقم 10 لا يمكن أن تكون الأقرب ١- مجاور إلى أي نقطة بيانات في مجموعة البيانات الاختيارية.

تشير نتائج التصنيف في المثال ٧-١ والخاصة عندما $k=3$، بالمقارنة مع نتائج التصنيف لـ $k=1$، إلا أن اختيار قيمة k يلعب دوراً هاماً في تحديد الفئة الهدف لنقطة البيانات. في المثال ٧-١، $k=1$ تعطى أداء أفضل من تصنيف $k=3$، وفي بعض الأمثلة أو التطبيقات الأخرى، إذا كانت k صغيرة جداً، على سبيل المثال، $k=1$، فإن أقرب ١- مجاور لنقطة البيانات x قد يكون نقطة بيانات شاذة أو القيمة التي تأتي من ضوضاء (*noise*) في مجموعة البيانات التدريبية. بجعل x تأخذ الفئة الهدف لهذا المجاور، لا يعطي المخرجات التي تعكس أنماط البيانات في مجموعة البيانات. إذا كانت k كبير جداً، قد تشمل مجموعة أقرب k- مجاور نقاط بيانات تقع بعيداً، والتي ليست حتى مشابهة لـ x. إن السماح لنقاط بيانات مختلفة باختيار فئة الهدف لـ x على أنها مجاورات لها يبدو أمراً غير عقلاني.

طريقة التعنقد المراقب في الجزء التالي تستخدم مصنِّف أقرب k- مجاور عن طريق تحديد عناقيد (*Clusters*) بيانات مماثلة أولاً ثم استخدام بيانات العناقيد هذه لتصنيف نقطة بيانات. وحيث إن بيانات العناقيد تعطي صورة أكثر تماسكاً عن مجموعة البيانات التدريبية من نقاط البيانات الفردية، فإن تصنيف نقطة بيانات ما على أساس عناقيد البيانات المجاورة لها وفئات الهدف الخاصة بها من المتوقع أن يعُطي أداء تصنيفي أكثر قوة من طريقة مصنف أقرب k- مجاورة، التي تعتمد على نقاط البيانات الفردية.

٧-٢ التعنقد المراقب (Supervised Clustering):

لقد تمّ تطوير خوارزمية التعنقد المراقب، وتمّ تطبيقها للكشف عن الهجمات عبر الإنترنت (*cyber attacks*) لتصنيف أنشطة طبيعة البيانات المرصودة والخاصة بالحاسوب والشبكات إلى وحدة من فئات الهدف: هجمات وأنشطة استخدام عادية (*Li and Ye, 2002, 2005, 2006; Ye, 2008, Ye and Li, 2002*). يمكن أن يتم تطبيق الخوارزمية أيضاً على مشاكل تصنيف أخرى.

للكشف عن الهجمات عبر الإنترنت، تحتوي البيانات التدريبية على كميات كبيرة من البيانات الحاسوبية وبيانات الشبكات لتعلم أنماط بيانات خاصة بالهجمات (*attacks*) وأنشطة الاستخدام العادي (*normal use activities*). بالإضافة إلى ذلك، يتم إضافة

المزيد من البيانات التدريبية مع مرور الوقت لتحديث أنماط البيانات الخاصة بالهجمات وأنشطة الاستخدام العادي. ومن ثم، يتطلب الأمر خوارزمية تعلم قابلة للتطوير المتزايد والقياس، بحيث يتم صيانة على أنماط البيانات الخاصة بالهجمات وأنشطة الاستخدام العادي، وتحديثها تدريجياً مع إضافة كل بيانات مرصودة جديدة بدلاً من معالجة كافة البيانات المرصودة في مجموعة البيانات التدريبية دفعةً واحدةً. وقد تمْ تطوير خوارزمية التعنقد المراقب باعتبارها خوارزمية تعلم قابلة للتطوير المتزايد والقياس لتعلم وتحديث أنماط البيانات لغرض التصنيف.

خلال عملية التدريب، فإن خوارزمية التعنقد المراقب تأخذ نقاط البيانات في مجموعة البيانات التدريبية واحدة تلو الأخرى لتجميعها في عناقيد من نقاط البيانات المتشابهة على أساس قيم متغيرات الخاصية، وقيم متغير الهدف الخاصة بها. يتم البدء بأول نقطة بيانات في مجموعة البيانات التدريبية، وجعل العنقود الأول يحتوى على نقطة البيانات هذه، ومن ثمّ أخذ فئة الهدف الخاصة بنقطة البيانات كفئة هدف لعنقود لبيانات. وعند أخذ نقطة البيانات الثانية في مجموعة البيانات التدريبية، نريد أن نجعل نقطة البيانات هذه تنضم إلى العنقود الأقرب الذي فئة هدفه نفس فئة هدف نقطة البيانات هذه. في خوارزمية التعنقد المراقب، نستخدم المتجه المتوسط (*mean vector*) لجميع نقاط البيانات في عنقود بيانات ما، على أنه المركز المتوسط (*centroid*) لعنقود البيانات الذي يتم استخدامه لتمثيل موقع عنقود البيانات، وحساب مسافة نقطة البيانات من هذا العنقود. إن عملية التعنقد (*clustering*) لا تستند فقط إلى قيم متغيرات الخاصية لقياس المسافة من نقطة البيانات إلى عنقود البيانات، ولكن أيضاً على الفئات الهدف لنقطة البيانات وعنقود البيانات لجعل نقطة البيانات تنضم إلى عنقود البيانات الذي له الفئة الهدف نفسه. جميع نقاط البيانات في نفس العنقود يكون لها نفس الفئة الهدف، والتي هي أيضاً الفئة الهدف للعنقود. ولأن الخوارزمية تستخدم الفئة الهدف لتوجيه أو للإشراف على تعنقد نقاط البيانات، فهي تُسمى خوارزمية التعنقد المراقب (*supervised clustering*).

لنفترض أن المسافة كبيرة من نقطة البيانات الأولى ونقطة البيانات الثانية في مجموعة البيانات التدريبية، ولكن نقطة البيانات الثانية لها نفس الفئة الهدف الخاصة بالعنقود الأول الذي يحتوي على نقطة البيانات الأولى، فإنه لا يزال على نقطة البيانات الثانية أن تنضم لهذا العنقود، لأنه هو عنقود البيانات الوحيد حتى الآن الذي لديه نفس الفئة الهدف. ومن ثم، فإن نتائج التعنقد تعتمد على الترتيب الذي تؤخذ به نقاط البيانات من مجموعة

البيانات التدريبية، مما يتسبب في مشكلة يطلق عليها التحيز المحلي لترتيب المدخلات (*local bias of the input order*). لمعالجة هذه المشكلة، فإن خوارزمية التعنقد المراقب تقوم بتجهيز عنقود بيانات مبدئي لكل فئة هدف. ولكل فئة هدف، يتم أولاً احتساب المركز المتوسط لجميع نقاط البيانات ذات الفئة الهدف في مجموعة البيانات التدريبية باستخدام المتجه المتوسط لنقاط البيانات. ثم يتم تجهيز عنقود مبدئي للفئة الهدف ليكون فيه المتجه المتوسط هو المركز المتوسط للعنقود والفئة الهدف، مما يعني الخروج بفئة هدف مختلفة عن أي فئة من فئات الهدف لنقاط البيانات في مجموعة البيانات التدريبية. على سبيل المثال، إذا كان لدينا فئتان من الفئات الهدف: T_1 وT_2 في البيانات التدريبية، يكون هناك عنقودان مبدئيان. العنقود المبدئي الأول يكون له المتجه المتوسط لنقاط البيانات T_1 كمركز متوسط (*centroid*). العنقود المبدئي الآخر يكون له المتجه المتوسط لنقاط البيانات لـ T_2 كمركز متوسط. يتم إسناد كل من العنقودين المبدئيين لفئة هدف، على سبيل المثال، T_3، والذي يختلف عن T_1 وT_2.

ولأن عناقيد البيانات الأولية هذه لا تحتوي على نقاط بيانات فردية، فإنه يُطلق عليها العناقيد الوهمية (*dummy clusters*). جميع العناقيد الوهمية تحتوى على فئة هدف تختلف عن أي فئة من الفئات الهدف في مجموعة البيانات التدريبية. تتطلب خوارزمية التعنقد المراقب من كل نقطة بيانات أن تقوم بتشكيل عنقود خاص بها، إذا ما كان عنقود البيانات الأقرب هو عنقود وهمي. مع العناقيد الوهمية، فنقطة البيانات الأولى من مجموعة البيانات التدريبية، تشكل عنقوداً جديداً لأنه لا يوجد إلا عناقيد وهمية فقط في البداية، والعنقود الأقرب إلى نقطة البيانات هذه هو عنقود وهمي.

إذا كانت نقطة البيانات الثانية لها نفس الفئة الهدف لنقطة البيانات الأولى، ولكنها تقع بعيداً عن نقطة البيانات الأولى، فمن الأرجح أن العنقود الوهمي يكون أقرب عنقود لنقطة البيانات الثانية من عنقود البيانات الذي يحتوي على نقطة البيانات الأولى. وهذا يجعل نقطة البيانات الثانية تشكل عنقوداً خاص بها، بدلاً من الانضمام إلى العنقود المحتوي على نقطة البيانات الأولى، ومن ثم فإن هذا يعالج مشكلة التحيز المحلي بسبب ترتيب المدخلات الخاصة بنقاط البيانات التدريبية.

خلال مرحلة الاختبار، تقوم خوارزمية التعنقد المراقب بتطبيق مصنِّف أقرب k- مجاور على عناقيد البيانات التي تمّ الحصول عليها من المرحلة التدريبية (أو الاستكشافية) من

خلال تحديد أقرب k- عنقود مجاور لنقطة البيانات المراد تصنيفها، ومن ثمّ جعل أقرب k- عنقود بيانات تصوت بالأغلبية لغرض تحديد الفئة الهدف لنقطة البيانات.

يوضح الجدول ٧-٣ الخطوات الخاصة بخوارزمية التعنقد المراقب. يتم استخدام الرموز التالية في وصف الخوارزمية:

$x_i=(x_{i,1},\ldots,x_{i,p},y_i)$: عبارة عن نقطة بيانات في مجموعة البيانات التدريبية بقيمة معروفة لـ y_i، لكل $i=1,\ldots,n$

$x=(x_i,\ldots,x_p,y)$: عبارة عن نقطة البيانات اختبارية وبقيمة لـ y يتم تحديدها لاحقاً

T_j : تمثل فئة الهدف رقم $j=1,\ldots,s$

C : تمثل عنقود بيانات

n_c : تمثل عدد نقاط البيانات في عنقود البيانات C

$\overline{x}_C$: تمثل المركز المتوسط لعنقود البيانات C والذي يمثل المتجه المتوسط لجميع نقاط البيانات في C

في الخطوة ٤ من المرحلة التدريبية (أو الاستكشافية)، بعد أن تنضم نقطة البيانات x_i إلى عنقود البيانات C، يتم تحديث المركز المتوسط لعنقود البيانات C تدريجياً لينتج $\overline{x}_C(t+1)$ (المركز المتوسط الذي تمّ تحديثه) باستخدام x_i، $\overline{x}_C(t)$ (المركز المتوسط الحالي للعنقود)، و$n_C(t)$ (العدد الحالي لنقاط البيانات في C):

$$\overline{x_C}(t+1) = \begin{bmatrix} \dfrac{n_C(t)\overline{x_{C1}}(t) + x_{i,1}}{n_C(t)+1} \\ \vdots \\ \dfrac{n_C(t)\overline{x_{Cp}}(t) + x_{i,p}}{n_C(t)+1} \end{bmatrix}. \qquad (١٤\text{-}٧)$$

خلال المرحلة التدريبية، يمكن إزالة العنقود الوهمي (*dummy cluster*) لفئة هدف معينة إذا تمّ إنشاء العديد من عناقيد البيانات لفئة الهدف هذه. وحيث إن المركز المتوسط (*centroid*) للعنقود الوهمي لفئة هدف معين هو المتجه المتوسط (*mean vector*) لجميع نقاط البيانات التدريبية ذات الفئة الهدف، فمن المرجح أن العنقود الوهمي للفئة الهدف هو العنقود الأقرب لنقطة البيانات. إزالة العنقود الوهمي للفئة الهدف يلغي هذا الاحتمال ويوقف إنشاء عنقود جديد لنقطة البيانات، لأن العنقود الوهمي للفئة الهدف هو العنقود الأقرب لنقطة البيانات.

الجدول (٧-٣)

خوارزمية التعنقد المراقب - (إنجليزي وعربي)

Step	Description
Training	
1	Set up *s* dummy clusters for *s* target classes, respectively, determine the centroid of each dummy cluster by computing the mean vector of all the data points in the training data set with the target class T_j, and assign T_{s+1} as the target class of each dummy cluster where $T_{s+1} \neq T_j$, $j = 1, \ldots, s$
2	FOR $i = 1$ to n
3	Compute the distance of x_i to each data cluster *C* including each dummy cluster, $d(x_i, \overline{x_C})$, using a measure of similarity
4	If the nearest cluster to the data point x_i has the same target class as that of the data point, let the data point join this cluster, and update the centroid of this cluster and the number of data points in this cluster
5	If the nearest cluster to the data point x_i has a different target class from that of the data point, form a new cluster containing this data point, use the attribute values of this data point as the centroid of this new cluster, let the number of data points in the cluster be 1, and assign the target class of the data point as the target class of the new cluster
Testing	
1	Compute the distance of the data point *x* to each data cluster *C* excluding each dummy cluster, $d(x, \overline{x_C})$
2	Let the *k*-nearest neighbor clusters of the data point vote for the target class of the data point

الخطوة	الوصف
	المرحلة التدريبية أو الاستكشافية (***Training***):
١	قم بتجهيز عدد s من العناقيد الوهمية (*dummy clusters*) لعدد s من الفئات الهدف (*target classes*)، على التوالي، ثم قم بتحديد المركز المتوسط (*centroid*) لكل عنقود وهمي عن طريق حساب المتجه المتوسط (*mean vector*) لجميع نقاط البيانات في مجموعة البيانات التدريبية والتي فئة هدفها تساوي T_j، ثم قم بإسناد فئة الهدف T_{s+1} كفئة هدف لكل عنقود وهمي بحيث أن $T_{s+1} \neq T_j, j=1, ..., s$
٢	كرر (*FOR*) ابتداء من $i=1$ إلى أن تصبح $i=n$
٣	احسب المسافة من x_i إلى كل عنقود بيانات C بما في ذلك كل عنقود وهمي، $d(x_i, \overline{x}_C)$، عن طريق استخدام مقياس للتشابه.
٤	إذا كان أقرب عنقود إلى نقطة البيانات x_i يحتوى على نفس فئة الهدف الموجودة في نقطة البيانات x_i، اجعل نقطة البيانات هذه تنظم إلى هذا العنقود، ثم قم بتحديث المركز المتوسط لهذا العنقود وتحديث عدد نقاط البيانات في هذا العنقود.
٥	إذا كان أقرب عنقود إلى نقطة البيانات x_i يحتوى على فئة هدف مختلفة عن تلك الموجودة في نقطة البيانات x_i قم بتشكيل أو إنشاء عنقود جديد يضم نقطة البيانات هذه، ثم قم باستخدام قيم متغيرات الخاصية لنقطة البيانات هذه كمركز متوسط لهذا العنقود الجديد، ثم اجعل عدد نقاط البيانات في العنقود يساوي ١، ثم قم بإسناد الفئة الهدف لنقطة البيانات كفئة هدف للعنقود الجديد.
	المرحلة الاختيارية (*Testing*):
١	احسب المسافة من نقطة البيانات x إلى كل عنقود بيانات C باستثناء كل عنقود وهمي $d(x_i, \overline{x}_C)$.
٢	اجعل أقرب k من العناقيد المجاورة لنقطة البيانات تقوم بالتصويت (*vote*) لغرض تحديد الفئة الهدف الخاصة بنقطة البيانات.

المثال (٧-٢):

استخدام خوارزمية التعنقد المراقب مع مقياس المسافة الإقليدية للاختلاف، ومصنِّف أقرب k- مجاور لتصنيف ما إذا كان نظام التصنيع معطلاً أم لا باستخدام مجموعة البيانات التدريبية في الجدول ٧-١، ومجموعة البيانات الاختيارية في الجدول ٧-٢. حيث تمّ شرح كلا الجدولين في المثال ٧-١.

في الخطوة ١ من المرحلة التدريبية، يتم تجهيز اثنين من العناقيد الوهمية C_1 وC_2 لاثنين من الفئات الهدف، $y=1$ و$y=0$، على التوالي:

$y_{C1}=2$ (تشير إلى أن C_1 هو عنقود وهمي بفئة هدف مختلفة عن فئتي هدف في مجموعات البيانات التدريبية ومجموعة البيانات الاختيارية).

$y_{C2}=2$ (تشير إلى أن C_2 هو عنقود وهمي)

$$\overline{x_{C_1}} = \begin{bmatrix} \frac{1+0+0+0+0+0+0+0+0}{9} \\ \frac{0+1+0+0+0+0+0+0+0}{9} \\ \frac{0+0+1+0+0+0+0+0+0}{9} \\ \frac{0+1+1+1+0+0+0+0+0}{9} \\ \frac{1+0+0+0+1+0+0+0+0}{9} \\ \frac{0+0+1+0+0+1+0+0+0}{9} \\ \frac{1+0+1+0+1+1+1+0+0}{9} \\ \frac{0+1+1+1+0+0+0+1+0}{9} \\ \frac{1+0+0+0+1+0+0+0+1}{9} \end{bmatrix} = \begin{bmatrix} 0.11 \\ 0.11 \\ 0.11 \\ 0.33 \\ 0.22 \\ 0.22 \\ 0.56 \\ 0.44 \\ 0.33 \end{bmatrix}$$

$$\overline{x_{C_2}} = \begin{bmatrix} \frac{0}{1} \\ \frac{0}{1} \\ \frac{0}{1} \\ \frac{0}{1} \\ \frac{0}{1} \\ \frac{0}{1} \\ \frac{0}{1} \\ \frac{0}{1} \\ \frac{0}{1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$n_{C_1} = 9$$

$$n_{C_2} = 1.$$

في الخطوة ٢ من المرحلة التدريبية، يتم البدء بمعالجة أول نقطة بيانات x_1 في مجموعة البيانات التدريبية:

$$x_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \quad y = 1.$$

في الخطوة ٣ من المرحلة التدريبية، يتم حساب المسافة الإقليدية من x_1 إلى كل من العناقيد الحالية C_1 وC_2:

$$d(x_1, \overline{x_{C_1}}) = \sqrt{\begin{matrix}(1-0.11)^2 + (0-0.11)^2 + (0-0.11)^2 + (0-0.33)^2 + (1-0.22)^2 \\ +(0-0.22)^2 + (1-0.56)^2 + (0-0.44)^2 + (1-0.33)^2\end{matrix}} = 1.56$$

$$d(x_1, \overline{x_{C_2}}) = \sqrt{\begin{matrix}(1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 \\ +(0-0)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2\end{matrix}} = 2$$

وحيث إن C_1 هو العنقود الأقرب إلى x_1 وله فئة هدف مختلفة عن تلك الخاصة ب x_1 يتم تنفيذ الخطوة ٥ من المرحلة التدريبية بتشكيل أو إنشاء عنقود بيانات جديد C_3 الذي يحتوي على x_1 :

$$y_{C_3} = 1$$

$$\overline{x_{C_3}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$n_{C_3} = 1.$$

بالعودة إلى الخطوة ٢ من المرحلة التدريبية، يتم البدء بمعالجة نقطة البيانات الثانية x_2 في مجموعة البيانات التدريبية:

$$x_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad y = 1.$$

في الخطوة ٣ من المرحلة التدريبية، يتم حساب المسافة الإقليدية من x_2 إلى كل من العناقيد الحالية C_1، C_2، وC_3:

$$d(x_2, \overline{x_{C_1}}) = \sqrt{\begin{matrix}(0-0.11)^2 + (1-0.11)^2 + (0-0.11)^2 + (1-0.33)^2 + (0-0.22)^2 \\ +(0-0.22)^2 + (0-0.56)^2 + (1-0.44)^2 + (0-0.33)^2\end{matrix}} = 1.44$$

$$d(x_2, \overline{x_{C_2}}) = \sqrt{\begin{matrix}(0-0)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 \\ +(0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2\end{matrix}} = 1.73$$

$$d(x_2, \overline{x_{C_3}}) = \sqrt{\begin{matrix}(0-1)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2 \\ +(0-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2\end{matrix}} = 2.65.$$

حيث إن C_1 هو العنقود الأقرب إلى x_2 وله فئة هدف مختلفة عن تلك التي لدى x_2 يتم تنفيذ الخطوة ٥ من المرحلة التدريبية بتشكيل أو إنشاء عنقود بيانات جديد C_4 الذي يحتوي على x_2:

$$y_{C_4} = 1$$

$$\overline{x_{C_4}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$n_{C_4} = 1.$$

بالعودة إلى الخطوة ٢ من المرحلة التدريبية، يتم البدء بمعالجة نقطة البيانات الثالثة x_3 في مجموعة البيانات التدريبية:

$$x_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \quad y = 1.$$

في الخطوة ٣ من المرحلة التدريبية، يتم حساب المسافة الإقليدية من x_3 إلى كل من العناقيد الحالية C_1، C_2، C_3، وC_4:

$$d\left(x_3, \overline{x_{C_1}}\right) = \sqrt{\begin{matrix}(0-0.11)^2 + (0-0.11)^2 + (1-0.11)^2 + (1-0.33)^2 + (0-0.22)^2 \\ +(1-0.22)^2 + (1-0.56)^2 + (1-0.44)^2 + (0-0.33)^2\end{matrix}} = 1.59$$

$$d(x_3, \overline{x_{C_2}}) = \sqrt{\begin{matrix}(0-0)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 \\ +(1-0)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2\end{matrix}} = 2.24$$

$$d(x_3, \overline{x_{C_3}}) = \sqrt{\begin{matrix}(0-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 \\ +(1-0)^2 + (1-1)^2 + (1-0)^2 + (0-1)^2\end{matrix}} = 2.45$$

$$d(x_3, \overline{x_{C_4}}) = \sqrt{\begin{matrix}(0-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2 \\ +(1-0)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2\end{matrix}} = 2.$$

حيث إن C_1 هو العنقود الأقرب إلى x_3، وله فئة هدف مختلفة عن تلك التي لدى x_3 يتم تنفيذ الخطوة ٥ من المرحلة التدريبية بتشكيل أو إنشاء عنقود بيانات جديد C_5 الذي يحتوي على x_3:

$$y_{C_5} = 1$$

$$\overline{x_{C_5}} = \begin{bmatrix}0\\0\\1\\1\\0\\1\\1\\1\\0\end{bmatrix}$$

$$n_{C_5} = 1.$$

بالعودة إلى الخطوة ٢ من المرحلة التدريبية، يتم معالجة نقطة البيانات الرابعة x_4 في مجموعة البيانات التدريبية:

$$x_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad y = 1.$$

في الخطوة ٣ من المرحلة التدريبية، يتم حساب المسافة الإقليدية من x_4 إلى كل من العناقيد الحالية C_1، C_2، C_3، C_4 وC_5:

$$d(x_4, \overline{x_{C_1}}) = \sqrt{\begin{matrix}(0-0.11)^2 + (0-0.11)^2 + (0-0.11)^2 + (1-0.33)^2 + (0-0.22)^2 \\ +(0-0.22)^2 + (0-0.56)^2 + (1-0.44)^2 + (0-0.33)^2\end{matrix}} = 1.14$$

$$d(x_4, \overline{x_{C_2}}) = \sqrt{\begin{matrix}(0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 \\ +(0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2\end{matrix}} = 1.41$$

$$d(x_4, \overline{x_{C_3}}) = \sqrt{\begin{matrix}(0-1)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2 \\ +(0-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2\end{matrix}} = 2.24$$

$$d(x_4, \overline{x_{C_4}}) = \sqrt{\begin{matrix}(0-0)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 \\ +(0-0)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2\end{matrix}} = 1$$

$$d(x_4, \overline{x_{C_5}}) = \sqrt{\begin{matrix}(0-0)^2 + (0-0)^2 + (0-1)^2 + (1-1)^2 + (0-0)^2 \\ +(0-1)^2 + (0-1)^2 + (1-1)^2 + (0-0)^2\end{matrix}} = 1.73.$$

حيث إن C_4 هو العنقود الأقرب إلى x_4 وله الفئة الهدف نفسها كما في x_4 يتم تنفيذ الخطوة ٤ من المرحلة التدريبية لإضافة x إلى العنقود C_4، والذي سيتم تحديثه لاحقاً:

$$y_{C_4} = 1$$

$$\overline{x_{C_4}} = \begin{bmatrix} \frac{0+0}{2} \\ \frac{1+0}{2} \\ \frac{0+0}{2} \\ \frac{1+1}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{1+1}{2} \\ \frac{0+0}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$n_{C_4} = 2.$$

تستمر المرحلة التدريبية أو الاستكشافية مع نقاط البيانات المتبقية x_5 x_6 x_7 x_8 وx_9 وتنتج العناقيد النهائية $C1$، $C2$، و $C3=\{x_1, x_5\}$، $C4=\{x_2, x_4\}$، $C5=\{x_3\}$، $C6=\{x_6\}$، $C7=\{x_7\}$، $C8=\{x_8\}$، $C9=\{x_9\}$، $C10=\{x_{10}\}$:

$$y_{C_1} = 2$$

$$\overline{x_{C_1}} = \begin{bmatrix} 0.11 \\ 0.11 \\ 0.11 \\ 0.33 \\ 0.22 \\ 0.22 \\ 0.56 \\ 0.44 \\ 0.33 \end{bmatrix}$$

$$n_{C_1} = 9$$

$$y_{C_2} = 2$$

$$\overline{x_{C_2}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$n_{C_2} = 1$$

$$y_{C_3} = 1$$

$$\overline{x_{C_3}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$n_{C_3} = 1$$

$$y_{C_4} = 1$$

$$\overline{x_{C_4}} = \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$n_{C_4} = 2$$

$$y_{C_5} = 1$$

$$\overline{x_{C_5}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

$$n_{C_5} = 1$$

$$y_{C_6} = 1$$

$$\overline{x_{C_6}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$n_{C_6} = 1$$

$$y_{C_7} = 1$$

$$\overline{x_{C_7}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$n_{C_7} = 1$$

$$y_{C_8} = 1$$

$$\overline{x_{C_8}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$n_{C_8} = 1$$

$$y_{C_9} = 1$$

$$\overline{x_{C_9}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$$n_{C_9} = 1$$

$$y_{C_{10}} = 0$$

$$\overline{x_{C_{10}}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$n_{C_{10}} = 1.$$

في مرحلة الاختبار، أول نقطة بيانات في مجموعة البيانات الاختبارية،

$$x = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

لها المسافات الإقليدية 1.73، 2.06، 2.45، 2.65، 2.45، 2.45، 2.45، و2.65 إلى العناقيد غير الوهمية C_3 ، C_4، C_5، C_6، C_7، C_8، C_9، وC_{10} على التوالي.

ومن ثم، فإن العنقود $C3$ هو المجاور الأقرب لـ x والفئة الهدف لـ x يتم إسنادها لتكون 1. العناقيد الأقرب لمجموعات نقاط البيانات المتبقية من 2 إلى 16 في مجموعة البيانات الاختيارية هي:

C_5، C_3،C_3 ، C_3، C_5، C_4، C_3/C_5، C_4 ، $C_3/C_6/C_{10}$، C_5، C_3، C_5 ، C_5، C_5، وC_3.

بالنسبة لنقطة البيانات 8، هناك تعادل بين C_3، وC_5 لغرض تحديد العنقود الأقرب. وحيث إن كلاً من C_3، وC_5 لهما الفئة الهدف 1، يتم إسناد فئة الهدف ١ لنقطة البيانات 8. بالنسبة لنقطة البيانات 10، هناك أيضاً تعادل بين C_3، C_6 ، C_{10}، لغرض تحديد أقرب عنقود. وحيث إن الغالبية (العنقودان C_3 وC_6) من العناقيد الثلاثة المتعادلة لها الفئة الهدف 1، يتم إسناد الفئة الهدف 1 إلى نقطة البيانات 10. ومن ثم، يتم إسناد كافة نقاط البيانات في مجموعة البيانات الاختيارية للفئة الهدف 1والتي صُنفت بشكل صحيح كما هو مبين في الجدول ٢-٢.

٧-٣ البرمجيات والتطبيقات (Software and Applications):

يمكن تطبيق مصنِّف أقرب k-مجاور وخوارزمية التعنقد المراقب بسهولة باستخدام برمجيات حاسوبية. ويمكن الاطلاع على تطبيق خوارزمية التعنقد المراقب لكشف الهجمات على الإنترنت في (*Li and Ye, 2002, 2005, 2006*)، وفي (*Ye, 2008*)، وفي (*Ye and Li, 2002*)

التمارين (Exercises):

٧-١ في مجموعة البيانات الخاصة بالحلقات الدائرية في مكوك الفضاء في الجدول ١-٢، المتغير الهدف هو عدد الحلقات الدائرية ذات الأحمال الثقيلة *(number of O-rings with Stress)*، له ثلاث قيم: 0، 1، و2. اعتبر هذه القيم الثلاث كقيم نوعية، في حين أن درجة حرارة الإطلاق *(Launch – Temperature)*، وضغط فحص التسرب *(Leak – check pressure)* هما متغيرات الخاصية، والحالات بالأرقام ١٣-٢٣ كبيانات تدريبية، والحالات بالأرقام ١-١٢ كبيانات اختبارية، والمسافة الإقليدية كمقياس للاختلاف. قم ببناء مُصنِّف أقرب ١ - مجاور، ومُصنِّف أقرب ٣- مجاور، ثم قم بفحص ومقارنة أدائهما التصنيفي.

٧-٢ أعد عمل التمرين ٧-١ باستخدام متغيرات الخاصية المطبعة من طريقة التطبيع في المعادلة ٧-٣.

٧-٣ أعد عمل التمرين ٧-١ باستخدام متغيرات الخاصية المطبعة من طريقة التطبيع في المعادلة ٧-٤.

٧-٤ باستخدام نفس مجموعتي البيانات التدريبية والبيانات الاختيارية في التمرين ٧-١، ومقياس تشابه جيب التمام (جتا) قم بإنشاء مصنِّف أقرب ١- مجاور، وإنشاء مصنِّف أقرب ٣- مجاور، ثم قم بفحص ومقارنة أدائهما التصنيفي.

٧-٥ باستخدام نفس مجموعتي البيانات التدريبية والبيانات الاختيارية في التمرين ٧-١، وخوارزمية التعنقد المراقب، ومقياس المسافة الإقليدية للاختلاف، قم ببناء مصنِّف أقرب ١- عنقود مجاور، وبناء مصنِّف أقرب ٣ - عنقود مجاور، ثم قم بفحص ومقارنة أدائهما التصنيفي.

٧-٦ أعد عمل التمرين ٧-٥ باستخدام متغيرات الخاصية المطبعة من طريقة التطبيع في المعادلة ٧-٣.

٧-٧ أعد عمل التمرين ٧-٥ باستخدام متغيرات الخاصية المطبعة من طريقة التطبيع في المعادلة ٧-٤.

٧-٨ باستخدام نفس مجموعتي البيانات التدريبية والبيانات الاختيارية في التمرين ٧-١، وخوارزمية التعنقد المراقب، ومقياس تشابه جيب التمام (جتا)، قم ببناء مصنّف أقرب ١- عنقود مجاور، وبناء مصنّف أقرب ٣ - عنقود مجاور، ثم قم بفحص ومقارنة أدائهما التصنيفي.

الجزء الثالث

خوارزميات لاستكشاف أنماط العنقود والاقتران

Algorithms for Mining Cluster and Association Patterns

٨- التعنقد الهرمي
Hierarchical Clustering

ينتج عن التعنقد الهرمي (*Hierarchical clustering*) مجموعات من سجلات البيانات المتشابهة على مستويات مختلفة من التشابه. يقدم هذا الفصل إجراء من أسفل إلى أعلى من التعنقد الهرمي، يُسمَى التعنقد الهرمي المحتشد (*agglomerative hierarchical clustering*). وترد قائمة من حزم البرمجيات التي تدعم التعنقد الهرمي. ويتم إعطاء بعض التطبيقات للتعنقد الهرمي مع مراجعتها.

٨-١ إجراء التعنقد الهرمي المحتشد
(Procedure of Agglomerative Hierarchical Clustering):

إذا كان لدينا عدد من سجلات البيانات في مجموعة البيانات، فإن استخدام خوارزمية التعنقد الهرمي المحتشد ينتج عنه عناقيد من سجلات البيانات المتشابهة حسب الخطوات التالية:

١. ابدأ بمجموعة عناقيد، كل منها يحتوي على سجل بيانات واحد.
٢. قم بدمج أقرب عنقودين لبعضهما لتشكيل عنقود جديد يستبدل العنقودين الأصليين ويحتوي على سجلات بيانات من العنقودين الأصليين.
٣. كرِّر الخطوة ٢ حتى يكون هناك عنقود واحد فقط يحتوي على كافة سجلات البيانات.

الجزء التالي يوضح استخدام طرق عدة لتحديد أقرب عنقودين في الخطوة ٢.

٢-٨ طرق تحديد المسافة بين عنقودين
(Methods of Determining the Distance between Two Clusters):

من أجل تحديد أقرب عنقودين في الخطوة ٢، نحتاج إلى طريقة لحساب المسافة بين العنقودين. يوجد عدد من الطرق والأساليب لتحديد المسافة بين العنقودين. يصف هذا الجزء أربعة طرق: طريقة الترابط المتوسط *(average linkage method)*، طريقة الترابط الأحادي (*single linkage*)، طريقة الترابط الكامل (*complete linkage*)، وطريقة المركز المتوسط (*centroid method*).

في طريقة الترابط المتوسط (*average linkage*)، فإن المسافة بين عنقودين (العنقود K ويرمز له، Ck، والعنقود L ويرمز له، CL) Dk,l، هي متوسط المسافات بين أزواج من سجلات البيانات (*pairs of data records*)، وكل زوج به سجل بيانات واحد من العنقود K وسجل بيانات آخر من العنقود L على النحو التالي:

$$D_{K,L} = \sum_{x_K \in C_K} \sum_{x_L \in C_L} \frac{d(x_K, x_L)}{n_K n_L} \qquad (١-٨)$$

$$x_K = \begin{bmatrix} x_{K,1} \\ \vdots \\ x_{K,p} \end{bmatrix} \qquad x_L = \begin{bmatrix} x_{L,1} \\ \vdots \\ x_{L,p} \end{bmatrix},$$

حيث إن:

x_K يدل على سجل بيانات في C_K

x_L يدل على سجل بيانات في C_L

n_K يدل على عدد سجلات البيانات في C_K

n_L يدل على عدد سجلات البيانات في C_L

$d(x_K, x_L)$ هي المسافة بين سجلي بيانات والتي يمكن حسابها باستخدام المسافة الإقليدية (*Euclidean distance*) التالية:

$$d(x_K, x_L) = \sum_{i=1}^{p} \left(x_{K,i} - x_{L,i}\right)^2 \qquad (٨-٢)$$

كما يمكن استخدام مقاييس تشابه/اختلاف بين نقطتي بيانات والتي تمّ توضيحها في الفصل ٧. وكما هو موضح في الفصل ٧، فإن تطبيع المتغيرات xp , ... ,$x1$ قد يكون ضرورياً قبل استخدام مقياس الاختلاف أو التشابه لحساب المسافة بين سجلي البيانات.

مثال (٨-١):

قم بحساب المسافة بين سجلي العنقودين التاليين باستخدام طريقة الترابط المتوسط والمسافة الإقليدية التربيعية لمجموعة من نقاط البيانات:

$$C_K = \{x_1, x_2, x_3\}$$

$$C_L = \{x_4, x_5\}$$

$$x_1 = \begin{bmatrix}1\\0\\0\\0\\1\\0\\1\\0\\1\end{bmatrix} \quad x_2 = \begin{bmatrix}0\\0\\0\\0\\1\\0\\1\\0\\1\end{bmatrix} \quad x_3 = \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\1\end{bmatrix} \quad x_4 = \begin{bmatrix}0\\0\\0\\0\\0\\1\\1\\0\\0\end{bmatrix} \quad x_5 = \begin{bmatrix}0\\0\\0\\0\\0\\0\\1\\0\\0\end{bmatrix}.$$

هناك ستة أزواج من سجلات البيانات بين C_K وC_L: (x_1 x_4)، (x_1 x_5)، (x_2 x_4)، (x_2 x_5)، (x_3 x_4)، (x_3 x_5)، ويتم حساب مسافتهم الإقليدية التربيعية كما يلي:

$$d(x_1, x_4) = \sum_{i=1}^{9} \left(x_{1,i} - x_{4,i}\right)^2$$

$$= (1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2$$

$$+ (1-1)^2 + (0-0)^2 + (1-0)^2 = 4$$

$$d(x_1, x_5) = \sum_{i=1}^{9} \left(x_{1,i} - x_{5,i}\right)^2$$

$$= (1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2$$

$$+ (1-1)^2 + (0-0)^2 + (1-0)^2 = 3$$

$$d(x_2, x_4) = \sum_{i=1}^{9} \left(x_{2,i} - x_{4,i}\right)^2$$

$$= (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2$$

$$+ (1-1)^2 + (0-0)^2 + (1-0)^2 = 3$$

$$d(x_2, x_5) = \sum_{i=1}^{9} \left(x_{2,i} - x_{5,i}\right)^2$$

$$= (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2$$

$$+ (1-1)^2 + (0-0)^2 + (1-0)^2 = 2$$

$$d(x_3, x_4) = \sum_{i=1}^{9} \left(x_{3,i} - x_{4,i}\right)^2$$

$$= (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2$$

$$+(0-1)^2+(0-0)^2+(1-0)^2=3$$

$$d(x_3,x_5)=\sum_{i=1}^{9}\left(x_{3,i}-x_{5,i}\right)^2$$

$$=(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2$$

$$+(0-1)^2+(0-0)^2+(1-0)^2=2$$

$$D_{K,L}=\sum_{x_K\in C_K}\sum_{x_L\in C_L}\frac{d(x_K,x_L)}{n_K n_L}$$

$$=\frac{4}{3\times 2}+\frac{3}{3\times 2}+\frac{3}{3\times 2}+\frac{2}{3\times 2}+\frac{3}{3\times 2}+\frac{2}{3\times 2}$$

$$=2.8333$$

في طريقة الترابط الأحادي (*single linkage*)، المسافة بين عنقودين تمثل المسافة الأقل بين سجل بيانات في عنقود واحد وسجل بيانات في العنقود الآخر:

$$D_{K,L}=\min\{d(x_K,x_L),x_k\in C_K,x_L\in C_L\}. \qquad (٣-٨)$$

باستخدام طريقة الترابط الأحادي، يتم حساب بالمسافة بين العنقودين C_K وC_L في المثال ٨-١ كما يلي:

$$D_{K,L}=\min\{d(x_K,x_L),x_k\in C_K,x_L\in C_L\}$$

$$=\min\{d(x_1,x_4),d(x_1,x_5),d(x_2,x_4),d(x_2,x_5),d(x_3,x_4),d(x_3,x_5)\}$$

$$=\min\{4,3,3,2,3,4\}=2.$$

في طريقة الترابط الكامل (*complete linkage*)، المسافة بين عنقودين تمثل المسافة الأكبر بين سجل بيانات في عنقود واحد وسجل بيانات في العنقود الآخر:

$$D_{K,L} = \max\{d(x_K, x_L), x_k \in C_K, x_L \in C_L\}. \qquad (٨-٤)$$

باستخدام طريقة الترابط الكامل، يتم حساب المسافة بين العنقودين C_K وC_L في المثال ٨-١ كما يلي:

$$D_{K,L} = \max\{d(x_K, x_L), x_k \in C_K, x_L \in C_L\}$$

$$= \max\{d(x_1, x_4), d(x_1, x_5), d(x_2, x_4), d(x_2, x_5), d(x_3, x_4), d(x_3, x_5)\}$$
$$= \max\{4, 3, 3, 2, 3, 4\} = 4.$$

في طريقة المركز المتوسط (*ceratoid*)، المسافة بين عنقودين تمثل المسافة بين المراكز المتوسطة للعناقيد، ويتم حساب المركز المتوسط لعنقود باستخدام المتجه المتوسط لجميع سجلات البيانات في العنقود، على النحو التالي:

$$D_{K,L} = d(\overline{x_K}, \overline{x_L}) \qquad (٨-٥)$$

$$\overline{x_K} = \begin{bmatrix} \frac{\sum_{k=1}^{n_K} x_{k,1}}{n_K} \\ \vdots \\ \frac{\sum_{k=1}^{n_K} x_{k,p}}{n_K} \end{bmatrix} \qquad \overline{x_L} = \begin{bmatrix} \frac{\sum_{l=1}^{n_L} x_{l,1}}{n_L} \\ \vdots \\ \frac{\sum_{l=1}^{n_L} x_{l,p}}{n_L} \end{bmatrix}. \qquad (٨-٦)$$

باستخدام طريقة ترابط المركز المتوسط (*centroid linkage method*) والمسافة الإقليدية التربيعية لنقاط البيانات، يتم حساب المسافة بين العنقودين C_K وC_L في المثال ٨-١ كما يلي:

$$\overline{x_K} = \begin{bmatrix} \frac{\sum_{k=1}^{n_K} x_{k,1}}{n_K} \\ \vdots \\ \frac{\sum_{k=1}^{n_K} x_{k,p}}{n_K} \end{bmatrix} = \begin{bmatrix} \frac{1+0+0}{3} \\ \frac{0+0+0}{3} \\ \frac{0+0+0}{3} \\ \frac{0+0+0}{3} \\ \frac{1+1+0}{3} \\ \frac{0+0+0}{3} \\ \frac{1+1+0}{3} \\ \frac{0+0+0}{3} \\ \frac{1+1+1}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ 0 \\ 0 \\ 0 \\ \frac{2}{3} \\ 0 \\ \frac{2}{3} \\ 0 \\ 1 \end{bmatrix}$$

$$\overline{x_L} = \begin{bmatrix} \frac{\sum_{l=1}^{n_L} x_{l,1}}{n_L} \\ \vdots \\ \frac{\sum_{l=1}^{n_L} x_{l,p}}{n_L} \end{bmatrix} = \begin{bmatrix} \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{1+0}{2} \\ \frac{1+1}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{2} \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{aligned} D_{K,L} = d(\overline{x_K}, \overline{x_L}) &= \left(\frac{1}{3} - 0\right)^2 + (1-0)^2 + (1-0)^2 \\ &+ (1-0)^2 + \left(\frac{2}{3} - 0\right)^2 + \left(0 - \frac{1}{2}\right)^2 + \left(\frac{2}{3} - 1\right)^2 \\ &+ (0-0)^2 + (1-0)^2 = 4.9167. \end{aligned}$$

يوجد طرق متنوعة لتحديد المسافة بين عنقودين، حيث إن استخدام هذه الطرق ينتج عنه مستويات مختلفة من التكلفة الحاسوبية اللازمة لإجراء العمليات الحسابية، وقد ينتج عنها نتائج تعنقد مختلفة. على سبيل المثال، فإن طريقة الترابط المتوسط، وطريقة الترابط

الأحادي، وطريقة الترابط الكامل تتطلب حساب المسافة بين كل زوج من نقاط البيانات من عنقودين. على الرغم من أن طريقة المركز المتوسط ليس لديها هذا المتطلب الحسابي، إلا أنه يجب على طريقة المركز المتوسط أن تحسب المركز المتوسط لكل عنقود جديد والمسافة من العنقود الجديد إلى العناقيد القائمة. إن طريقة الترابط المتوسط وطريقة المركز المتوسط تأخذ بعين الاعتبار وتتحكم بانتشار وتشتت نقاط البيانات في كل عنقود، في حين أن طريقة الترابط الأحادي وطريقة الترابط الكامل لا تضع أية قيود على شكل العنقود.

٨-٣ توضيح كيفية إجراء التعنقد الهرمي (Illustration of the Hierarchical Clustering Procedure):

يتم توضيح إجراء التعنقد الهرمي في المثال ٨-٢.

المثال (٨-٢):

قم بإجراء التعنقد الهرمي على بيانات اكتشاف أعطال النظام في الجدول ٨-١ باستخدام طريقة الترابط الأحادي.

الجدول (٨-١)

مجموعة البيانات الخاصة باكتشاف أعطال النظام مع تسع حالات من الأعطال الآلية الأحادية

رقم الحالة - Instance (الآلة المعطلة - Faulty Machine)	متغيرات الخاصية عن جودة وحدات المنتج Attribute Variables about Quality of Parts								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1 (M1)	1	0	0	0	1	0	1	0	1
2(M2)	0	1	0	1	0	0	0	1	0
3(M3)	0	0	1	1	0	1	1	1	0
4(M4)	0	0	0	1	0	0	0	1	0
5(M5)	0	0	0	0	1	0	1	0	1
6(M6)	0	0	0	0	0	1	1	0	0
7(M7)	0	0	0	0	0	0	1	0	0
8(M8)	0	0	0	0	0	0	0	1	0
9(M9)	0	0	0	0	0	0	0	0	1

يحتوي الجدول ٨-١ على مجموعة البيانات لاكتشاف أعطال النظام، بما في ذلك تسع حالات للأعطال الآلية الأحادية. ويتم استخدام متغيرات الخاصية التسعة الخاصة بجودة وحدات المنتج فقط في التعنقد الهرمي. وفيما يلي نسرد سجلات البيانات التسعة في مجموعة البيانات.

$$x_1 = \begin{bmatrix}1\\0\\0\\0\\1\\0\\1\\0\\1\end{bmatrix} \quad x_2 = \begin{bmatrix}0\\1\\0\\1\\0\\0\\0\\1\\0\end{bmatrix} \quad x_3 = \begin{bmatrix}0\\0\\1\\1\\0\\1\\1\\1\\0\end{bmatrix} \quad x_4 = \begin{bmatrix}0\\0\\0\\1\\0\\0\\0\\1\\0\end{bmatrix} \quad x_5 = \begin{bmatrix}0\\0\\0\\0\\1\\0\\1\\0\\1\end{bmatrix} \quad x_6 = \begin{bmatrix}0\\0\\0\\0\\0\\1\\1\\0\\0\end{bmatrix}$$

$$x_7 = \begin{bmatrix}0\\0\\0\\0\\0\\0\\1\\0\\0\end{bmatrix} \quad x_8 = \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\1\\0\end{bmatrix} \quad x_9 = \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\1\end{bmatrix}.$$

سوف تُظهِر نتائجُ التعنقد أياً من الأعطال الآلية الأحادية يوجد بها أعراض متشابهة تخص مشكلة جودة وحدات المنتج.

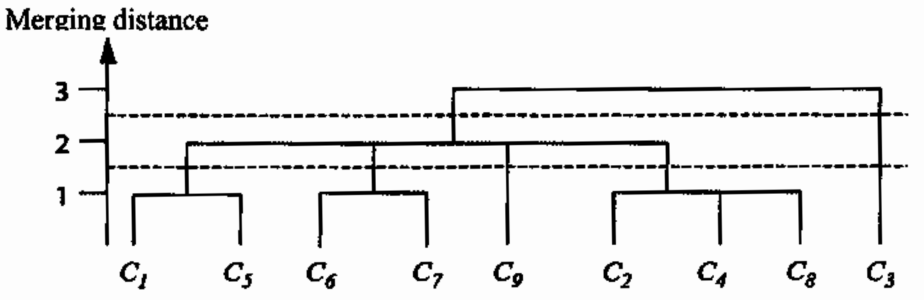
ويبين الشكل ٨-١ إجراء التعنقد الهرمي الذي يبدأ بالعناقيد التسعة التالية وبسجل بيانات واحد في كل عنقود:

$$C_1 = \{x_1\} \quad C_2 = \{x_2\} \quad C_3 = \{x_3\} \quad C_4 = \{x_4\} \quad C_5 = \{x_5\}$$

$$C_6 = \{x_6\} \quad C_7 = \{x_7\} \quad C_8 = \{x_8\} \quad C_9 = \{x_9\}.$$

الشكل (٨-١)

نتيجة التعنقد الهرمي لمجموعة بيانات اكتشاف أعطال النظام



الجدول (٨-٢)

المسافة لكل زوج من العناقيد: C_1، C_2، C_3، C_4، C_5، C_6، C_7، C_8، و C_9

$C_9 = \{x_9\}$	$C_8 = \{x_8\}$	$C_7 = \{x_7\}$	$C_6 = \{x_6\}$	$C_5 = \{x_5\}$	$C_4 = \{x_4\}$	$C_3 = \{x_3\}$	$C_2 = \{x_2\}$	$C_1 = \{x_1\}$	
3	5	3	4	1	6	7	7		$C_1=\{x_1\}$
4	2	4	5	6	1	4			$C_2=\{x_2\}$
6	4	4	6	6	3				$C_3=\{x_3\}$
3	1	4	4	5					$C_4=\{x_4\}$
2	4	2	3						$C_5=\{x_5\}$
3	3	1							$C_6=\{x_6\}$
2	2								$C_7=\{x_7\}$
2									$C_8=\{x_8\}$
									$C_9=\{x_9\}$

نظراً لأن كل عنقود يحتوي سجل بيانات واحد فقط، فإن المسافة بين عنقودين هي المسافة بين سجلي البيانات في العنقودين، على التوالي. يوضح الجدول ٨-٢ المسافة لكل زوج من سجلات البيانات، والتي تمثل أيضا المسافة لكل زوج من العناقيد.

هناك أربعة أزواج من العناقيد ينتج عنها أصغر مسافة بقيمة تساوي 1: (C_1, C_5)، (C_2, C_4)، (C_4, C_8)، و(C_6, C_7). نقوم بدمج (C_1, C_5) لتشكيل عنقود جديد $C_{1,5}$ ودمج (C_6, C_7) لتشكيل عنقود جديد $C_{6,7}$. وحيث يشترك العنقود C_4 في اثنين من أزواج العناقيد (C_2, C_4) و(C_4, C_8)، فيمكننا دمج زوج واحد فقط من العناقيد. نختار بشكل عشوائي أن ندمج (C_2, C_4) لتشكيل عنقود جديد $C_{2,4}$. ويبين الشكل ٨-١ هذه العناقيد الجديدة، في مجموعة جديدة من العناقيد $C_{1,5}$، $C_{2,4}$، C_3، $C_{6,7}$، C_8 و C_9.

يعطي الجدول ٨-٣ المسافة لكل زوج من العناقيد، $C_{1,5}$، $C_{2,4}$، C_3، $C_{6,7}$، C_8، وC_9، باستخدام طريقة الترابط الأحادي. على سبيل المثال، هناك أربعة أزواج من سجلات البيانات بين $C_{1,5}$، و$C_{2,4}$: $(x_1\ x_2)$، $(x_1\ x_4)$، $(x_5\ x_2)$، و$(x_5\ x_4)$، وبالمسافات التي بينهم 7، 6، 6، و 5، على التوالي، من الجدول ٨-٢. ومن ثم، فإن المسافة الأقل من بين هذه المسافات هي 5، والتي تُؤخَذ على أنها المسافة بين $C_{1,5}$ و $C_{2,4}$.

الجدول (٨-٣)

مسافة كل زوج من العناقيد: $C_{1,5}$، $C_{2,4}$، C_3، $C_{6,7}$، C_8 و C_9

$C_9 = \{x_9\}$	$C_8 = \{x_8\}$	$C_{6,7} = \{x_6, x_7\}$	$C_3 = \{x_3\}$	$C_{2,4} = \{x_2, x_4\}$	$C_{1,5} = \{x_1, x_5\}$	
2 = min {3, 2}	4 = min {5, 4}	2 = min {4, 3, 3, 2}	6 = min {7, 6}	5 = min {7, 6, 6, 5}		$C_{1,5} = \{x_1, x_5\}$
3 = min {4, 3}	1 = min {2, 1}	4 = min {5, 4, 4, 4}	3 = min {4, 3}			$C_{2,4} = \{x_2, x_4\}$
6 = min {6}	4 = min {4}	5 = min {6, 4}				$C_3 = \{x_3\}$
2 = min {3, 2}	2 = min {3, 2}					$C_{6,7} = \{x_6, x_7\}$
2 = min {2}						$C_8 = \{x_8\}$
						$C_9 = \{x_9\}$

الجدول (٨-٤)

مسافة كل زوج من العناقيد: $C_{1,5}$، $C_{2,4,8}$، C_3، $C_{6,7}$، و C_9

	$C_{1,5} = \{x_1, x_5\}$	$C_{2,4,8} = \{x_2, x_4, x_8\}$	$C_3 = \{x_3\}$	$C_{6,7} = \{x_6, x_7\}$	$C_9 = \{x_9\}$
$C_{1,5} = \{x_1, x_5\}$		4 = min {7, 6, 5, 6, 5, 4}	6 = min {7, 6}	2 = min {4, 3, 3, 2}	2 = min {3, 2}
$C_{2,4,8} = \{x_2, x_4, x_8\}$			3 = min {4, 3, 4}	2 = min {5, 4, 4, 4, 3, 2}	3 = min {4, 3, 2}
$C_3 = \{x_3\}$				4 = min {6, 4}	6 = min {6}
$C_{6,7} = \{x_6, x_7\}$					2 = min {3, 2}
$C_9 = \{x_9\}$					

إن أقرب زوج من العناقيد هو ($C_{2,4}$ ،C_8) بمسافة تساوي 1. دمج العنقودين $C_{2,4}$ C_8، ينتج عنقوداً جديداً هو $C_{2,4,8}$. ويكون مجموعة جديدة من العناقيد، $C_{1,5}$، $C_{2,4,8}$، C_3، $C_{6,7}$، وC_9

يعطي الجدول ٨-٤ المسافة لكل زوج من العناقيد، $C_{1,5}$، $C_{2,4,8}$، C_3، $C_{6,7}$، وC_9، باستخدام طريقة الترابط الأحادي. أربعة أزواج من العناقيد، ($C_{1,5}$ ،$C_{6,7}$) و($C_{1,5}$ ،C_9)، ($C_{2,4,8}$ ،$C_{6,7}$)، و($C_{6,7}$ ،C_9)، ينتج عنها أصغر مسافة وتساوي 2. حيث أن العناقيد الثلاثة $C_{1,5}$ ،$C_{6,7}$ ،C_9 تبعد نفس المسافة بعضها عن بعض، نقوم بدمج الثلاثة عناقيد معاً لتشكيل عنقود جديد، $C_{1,5,6,7,9}$. لا يتم دمج $C_{6,7}$ مع $C_{2,4,8}$ لأن $C_{6,7}$ قد تمّ دمجه مع $C_{1,5}$ وC_9. ويكون لدينا مجموعة جديدة من العناقيد، $C_{1,5,6,7,9}$، $C_{2,4,8}$، C_3.

يعطي الجدول ٨-٥ المسافة لكل زوج من العناقيد، $C_{1,5,6,7,9}$، $C_{2,4,8}$، C_3، وذلك باستخدام طريقة الترابط الأحادي. ينتج زوج العناقيد، ($C_{1,5,6,7,9}$, $C_{2,4,8}$)، أصغر مسافة وتساوي 2. دمج العناقيد، $C_{1,5,6,7,9}$ و$C_{2,4,8}$، يشكل عنقود جديد، $C_{1,2,4,5,6,7,8,9}$. ويكون لدينا مجموعة جديدة من العناقيد، $C_{1,2,5,4,5,6,7,8,9}$، وC_3 والتي لديها مسافة 3 ويتم دمجها في عنقود واحد، $C_{1,2,3,4,5,6,7,8,9}$.

الجدول (٨-٥)

مسافة كل زوج من العناقيد: $C_{1,5,6,7,9}$، $C_{2,4,8}$، و C_3

	$C_{1,5,6,7,9} = \{x_1, x_5, x_6, x_7, x_9\}$	$C_{2,4,8} = \{x_2, x_4, x_8\}$	$C_3 = \{x_3\}$
$C_{1,5,6,7,9} = \{x_1, x_5, x_6, x_7, x_9\}$		$2 = \min \{7, 6, 5, 6, 5, 4, 5, 4, 3,4, 4, 2, 4, 3, 2\}$	$4 = \min \{7, 6, 6, 4, 6\}$
$C_{2,4,8} = \{x_2, x_4, x_8\}$			$3 = \min \{4, 3, 4\}$
$C_3 = \{x_3\}$			

ويبين الشكل ٨-١ أيضاً مسافة الدمج، والتي تمثل المسافة بين عنقودين عندما يتم دمجهما معاً. تُسمى شجرة التعنقد الهرمي الموضحة في الشكل ٨-١ برسم الدندروقرام الهرمي (*dendrogram*).

يسمح التعنقد الهرمي بالحصول على مجموعات مختلفة من العناقيد من خلال وضع حدود (*thresholds*) مختلفة لحد مسافة الدمج لغرض وضع مستويات مختلفة من تشابه البيانات. على سبيل المثال، إذا وضعنا حد مسافة الدمج تساوي 1.5 كما هو موضح بالخط المقطّع في الشكل ٨-١، نحصل على العناقيد، $C_{1,5}$، $C_{6,7}$، C_9، $C_{2,4,8}$، C_3 والتي تعد عناقيد بيانات متشابهة نظراً لأن مسافة الدمج لكل عنقود هي أصغر من أو تساوي الحد 1.5. تشير هذه المجموعة من العناقيد إلى أي الأعطال الآلية تعطي أعراضاً متشابهةً لمشكلة جودة وحدات المنتج. على سبيل المثال، العنقود $C_{1,5}$، يشير إلى أن عطل الآلة الأولى $M1$ وعطل الآلة الخامسة $M5$ ينتجان أعراضاً متشابهةً لمشكلة جودة وحدات المنتج. يبين تدفق إنتاج وحدات المنتج في الشكل ١-١ أن وحدات المنتج تمر عبر الآلتين الأولى $M1$ والخامسة $M5$ على التوالي، ومن ثم هذا يُفسر سبب أن عطل الآلة الأولى $M1$ وعطل الآلة الخامسة $M5$ ينتجان أعراضاً متشابهة لمشكلة جودة وحدات المنتج. ومن ثم، فإن العناقيد التي تمّ الحصول عليها عن طريق تحديد حد مسافة الدمج إلى 1.5 تعطي نتيجة عنقودية ذات معنى والتي تكشف عن الهيكل المترابط للنظام. إذا وضعنا حد مسافة الدمج يساوي 2.5 كما هو موضح بخط مقطّع آخر في الشكل ٨-١، فإننا نحصل على مجموعة من العناقيد،

$C_{1,2,4,5,6,7,8,9}$ و C_3، والتي ليست بمستوى فائدة مجموعة العناقيد، $C_{1,5}$،$C_{6,7}$، C_9، $C_{2,4,8}$ وC_3 للكشف عن هيكل المستخدم.

يوضح هذا المثال أن الحصول على نتيجة استكشاف البيانات ليست نهاية عملية الاستكشاف. فمن الأهمية بمكان أن نتمكن من توضيح نتيجة استكشاف البيانات بطريقة ذات معنى في سياق المشكلة المبحوثة أو المستهدفة لجعل هذه النتيجة مفيدة في مجال ونطاق المشكلة. العديد من مجموعات البيانات في العالم الحقيقي لا تكون مصحوبة بمعرفة مسبقة للنظام الذي قام بتوليد هذه المجموعات من البيانات. ولذلك، بعد الحصول على نتيجة التعنقد الهرمي، فمن المهم دراسة مجموعات مختلفة من العناقيد على مستويات مختلفة من تشابه البيانات ومن ثم تحديد أي مجموعة من العناقيد يمكن تفسيرها بطريقة ذات معنى للمساعدة في الكشف عن النظام وتوليد معرفة مفيدة عن النظام.

٨-٤ الشجرة غير الرتيبة للتعنقد الهرمي (Nonmonotonic Tree of Hierarchical Clustering):

في الشكل ٨-١، لا تكون مسافة دمج عنقود جديد أصغر من مسافة دمج أي عنقود تمّ إنشاؤه قبل العنقود الجديد. وشجرة التعنقد الهرمي هذه تكون رتيبة (*monotonic*). على سبيل المثال، في الشكل ٨-١، مسافة دمج العنقود $C_{2,4}$، هي 1، وهي تساوي مسافة دمج $C_{2,4,8}$، ومسافة دمج $C_{1,2,4,5,6,7,8,9}$، هي 2، والتي هي أصغر من مسافة دمج $C_{2,4,8}$.

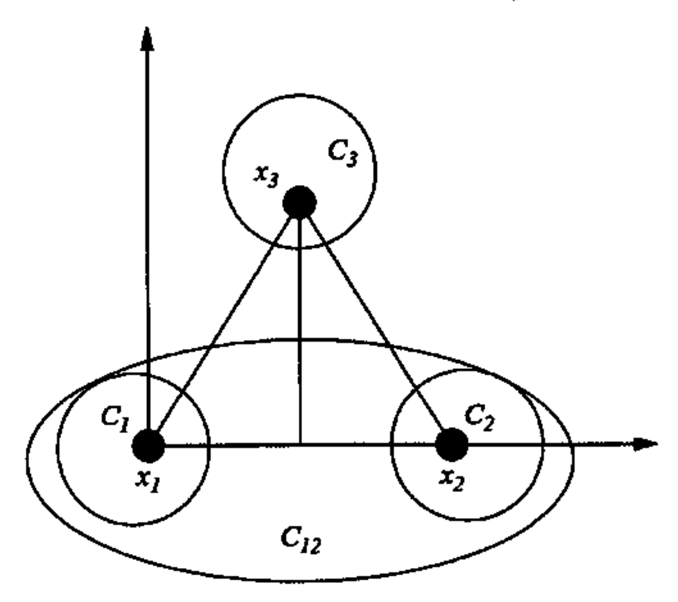
إنَّ طريقة ترابط المركز المتوسط يمكن أن تنتج شجرة غير رتيبة (*non monotonic tree*) والتي يمكن أن تكون فيها مسافة الدمج لعنقود جديد أصغر من مسافة الدمج لعنقود يتم إنشاؤه قبل العنقود الجديد. الشكل ٨-٢ يظهر ثلاث نقاط بيانات، x_1 x_2 وx_3 واللاتي باستخدامهن تقوم طريقة المركز المتوسط بإنتاج شجرة غير رتيبة للتعنقد الهرمي (*non monotonic tree of hierarchical clustering*). المسافة بين كل زوج من نقاط البيانات الثلاثة هي 2. نبدأ بالعناقيد الأولية الثلاثة، C_1، C_2، C_3، والمحتوية على ثلاث نقاط بيانات، x_1 x_2 وx_3 على التوالي. ونظراً لأن العناقيد الثلاثة لها المسافة نفسها بين بعضها البعض، فنختار بشكل عشوائي دمج C_1 وC_2 في عنقود جديد $C_{1,2}$. كما هو موضح في الشكل ٨-٢، فإن المسافة بين المركز المتوسط لــ $C_{1,2}$ و x_3 هي: $1.73 = \sqrt{2^2 - 1^2}$، والتي هي أصغر من مسافة دمج المساوية 2 لـ $C_{1,2}$، ومن ثم، عندما يتم

دمج $C_{1,2}$ مع C_3 بعد ذلك لإنتاج عنقود جديد $C_{1,2,3}$، تكون مسافة الدمج 1.73 لـ $C_{1,2,3}$ أصغر من مسافة الدمج 2 لـ $C_{1,2}$. الشكل ٨-٣ يوضح الشجرة غير الرتيبة للتعنقد الهرمي لنقاط البيانات الثلاثة هذه باستخدام طريقة المركز المتوسط.

طريقة الترابط الأحادي، التي تمّ استخدامها في المثال ٨-٢، تقوم بحساب المسافة بين عنقودين باستخدام أصغر مسافة بين نقطتي بيانات، نقطة بيانات واحدة في عنقود واحد، ونقطة بيانات أخرى في العنقود الآخر. تُستخدم أصغر مسافة بين نقطتي بيانات لتشكيل وإنشاء عنقود جديد. المسافة المستخدمة لتشكيل وإنشاء عنقود مسبقاً لا يمكن استخدامها مرة أخرى لتشكيل عنقود جديد لاحق، لأن المسافة تصبح بالفعل داخل عنقود وهناك حاجة إلى مسافة لنقطة بيانات خارج عنقود ما لتشكيل عنقود جديد في وقت لاحق. ومن ثم، فإن المسافة لتشكيل عنقود جديد في وقت لاحق يجب أن تأتي من مسافة لم تُستخدَم من قبل، والتي يجب أن تكون أكبر من أو تساوي مسافة تمّ اختيارها واستخدامها في وقت سابق. ومن ثم، فإن شجرة التعنقد الهرمي من طريقة الترابط الأحادي هي دائماً رتيبة.

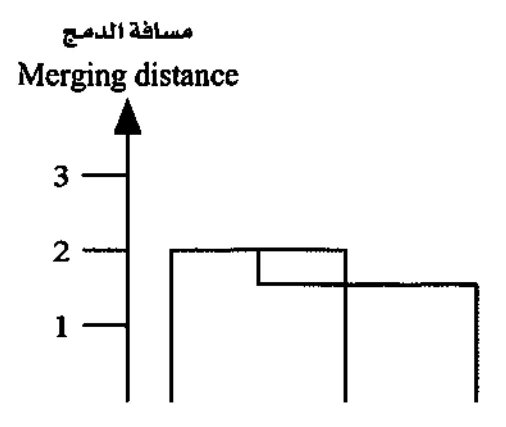
الشكل (٨-٢)

مثال على ثلاث نقاط بيانات والتي تنتج لها طريقة ترابط المركز المتوسط شجرة غير رئيسية للتعنقد الهرمي



الشكل (٨-٣)

الشجرة غير الرئيسية للتعنقد الهرمي لنقاط البيانات في الشكل (٨-٢)



٨-٥ البرمجيات والتطبيقات (Software and Applications):

يتم دعم التعنقد الهرمي بالعديد من الحزم البرمجية الإحصائية، بما في ذلك:

- *SAS (www.sas.com)*
- *SPSS (www.spss.com)*
- *STATISTICA (www.statistica.com)*
- *MATLAB ® (www.matworks.com)*

يمكن العثور على بعض تطبيقات التعنقد الهرمي في الأعمال التالية: (*Ye, 1997, 2003, Chapter 10; Ye and Salvendy, 1991, 1994; Ye and Zhao, 1996*). في العمل الذي أجراه يي وسالفيندي (*Ye and Salvendy, 1994*)، يتم استخدام التعنقد الهرمي للكشف عن التركيبة المعرفية للغة البرمجة سي (*C*) والموجودة لدى المبرمجين الخبراء والمبرمجين المبتدئين.

التمارين (Exercises):

٨-١ قُم بعمل تعنقد هرمي لـــ ٢٣ سجلاً من سجلات البيانات في مجموعات البيانات الدائرية في مكوك الفضاء الواردة في الجدول ١-٢. استخدم درجة حرارة الإطلاق (*Launch- Temperature*) وضغط التحقق من التسرب (*Leak- Check Pressure*) كمتغيرات الخاصية، وطريقة التطبيع في المعادلة ٧-٤ للحصول على قيم مطبّعة لدرجة حرارة الإطلاق وضغط التحقق من التسرب أيضاً، والمسافة الإقليدية لسجلات البيانات، وطريقة الترابط الأحادي.

٨-٢ كرِّر التمرين ٨-١ باستخدام طريقة الترابط الكامل.

٨-٣ كرِّر التمرين ٨-١ باستخدام مقياس تشابه جيب التمام (جتا) *(cosine similarity)* لحساب المسافة بين سجلات البيانات.

٨-٤ كرِّر التمرين ٨-٣ باستخدام طريقة الترابط الكامل.

٨-٥ ناقش فيما إذا كان ممكناً أو غير ممكن إنتاج شجرة غير رتيبة للتعنقد الهرمي باستخدام طريقة الترابط الكامل.

٨-٦ ناقش فيما إذا كان ممكناً أو غير ممكن إنتاج شجرة غير رتيبة للتعنقد الهرمي باستخدام طريقة الترابط المتوسط.

٩- التعنقد حول K- متوسط والتعنقد القائم على الكثافة
K–Means Clustering and Density–Based Clustering

يستعرض هذا الفصل خوارزميات التعنقد حول K- متوسط (*K- Means clustering*) والتعنقد القائم على الكثافة (*Density- Based Clustering*)، والتي ينتج عنها مجموعات غير هرمية من سجلات البيانات المتشابهة، باستخدام المركز المتوسط (*centroid*) والكثافة (*density*) لعنقود ما، على التوالي. وسيتم سرد قائمة بحزم البرمجيات التي تدعم خوارزميات التعنقد هذه. وسيتم سرد قائمة لبعض تطبيقات خوارزميات التعنقد مع مراجعها.

٩-١ التعنقد حول K- متوسط (***K*-Means Clustering**):

يرد في الجدول ٩-١ خطوات خوارزمية التعنقد حول K-متوسط. تبدأ خوارزمية التعنقد حول K-متوسط بقيمة معينة لـ K والقيم الأولية المُسنَده للمراكز المتوسطة والخاصة بعدد K من العناقيد. وتستمر الخوارزمية بجعل كل سجل من سجلات البيانات التي عددها n في مجموعة البيانات تنضم إلى أقرب عنقود لها وتحديث المراكز المتوسطة للعناقيد حتى لا تتغير قيم المراكز المتوسطة للعناقيد بعد ذلك، ونتيجةً لذلك لا ينتقل كل سجل بيانات من عنقوده الحالي إلى عنقود آخر. في الخطوة ٧ من الخوارزمية، إذا كان هناك أي تغيير على قيم المراكز المتوسطة للعناقيد في الخطوات من ٣ إلى ٦، فيتعين علينا معرفة ما إذا كان التغيير على قيم المراكز المتوسطة للعناقيد يتسبب في المزيد من التنقل لأي سجل بيانات من خلال العودة إلى الخطوة ٢.

لتحديد أقرب عنقود إلى سجل بيانات، فإن المسافة من سجل البيانات إلى عنقود البيانات تحتاج إلى أن يتم حسابها. وغالباً ما يتم استخدام المتجه المتوسط لسجلات البيانات في عنقود ما كمركز متوسط للعنقود. باستخدام مقياس للتشابه أو الاختلاف، نقوم بحساب المسافة من سجل البيانات إلى المركز المتوسط للعنقود لتمثل المسافة من سجل البيانات إلى العنقود. ويمكن الرجوع إلى فصل ٧ للحصول على وصف واف لمقاييس التشابه والاختلاف.

إحدى الطرق لإسناد قيم أولية للمراكز المتوسطة الخاصة بعدد K من العناقيد تكون باختيار عدد K من سجلات البيانات عشوائياً من مجموعة البيانات واستخدام سجلات البيانات هذه لبناء قيم المركز المتوسطة لـ K من العناقيد. على الرغم من أن هذه الطريقة تستخدم سجلات بيانات محددة لبناء قيم المراكز المتوسطة لـ K من العناقيد، إلا أن الـ K-عنقود لا يوجد بها سجل بيانات واحد في كل منها في البداية. هناك أيضاً طرق أخرى لإعطاء قيم أولية للمراكز المتوسطة الخاصة بـ K من العناقيد، مثل استخدام نتيجة التعنقد الهرمي للحصول على عدد K من العناقيد واستخدام المراكز المتوسطة لهذه العناقيد كمراكز متوسطة أولية الخاصة بـ K من العناقيد لغرض استخدامها في خوارزمية التعنقد حول K-متوسط.

بالنسبة إلى مجموعة بيانات كبيرة في الحجم، فإن شرط التوقف لتعليمة التكرار (*REPEAT-UNTIL*) في الخطوة رقم ٧ من الخوارزمية يمكن أن يتم تحقيقه، بحيث تتوقف تعليمة التكرار عندما يكون مقدار التغييرات للمراكز المتوسطة أقل من حد معين، على سبيل المثال، أقل من ٥٪ من سجلات البيانات التي تغير عناقيدها المحتوية لها.

الجدول (٩-١)

خوارزمية التعنقد حول K-متوسط - (إنجليزي وعربي)

Description
Set up the initial centroids of the K clusters
REPEAT
FOR $i = 1$ to n
Compute the distance of the data point $\mathbf{x}_i$ to each of the K clusters using a measure of similarity or dissimilarity
IF $\mathbf{x}_i$ is not in any cluster or its closest cluster is not its current cluster
Move $\mathbf{x}_i$ to its closest cluster and update the centroid of the cluster
UNTIL no change of centroid clusters occurs in Steps 3–6

الخطوة	الوصف
١	قم بتجهيز المراكز المتوسطة الأولية لعدد K من العناقيد.
٢	كرر *(REPEAT)*.
٣	كرر *(FOR)* ابتداء من $i=1$ إلى n.
٤	قم بحساب المسافة من سجل البيانات x_i إلى كل العناقيد التي عددها K باستخدام مقياس التشابه أو الاختلاف.
٥	إذا *(IF)* لم تكن x_i في أي عنقود أو أن أقرب عنقود لها ليس هو عنقودها الحالي.
٦	قم بنقل x_i إلى أقرب عنقود وقم بتحديث المركز المتوسط للعنقود.
٧	حتى *(UNTIL)* الوقت الذي لا يحدث به تغير في المراكز المتوسط للعناقيد في الخطوات ٣-٦.

تُقلل خوارزمية التعنقد حول K- متوسط من مجموع الأخطاء التربيعية (*sum of squared errors-SSE*) التالية أو المسافات بين سجلات البيانات والمراكز المتوسطة للعناقيد (*Ye, 2003, Chapter 10*):

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} d\left(x, \overline{x_{C_i}}\right)^2 . \qquad (١-٩)$$

في المعادلة ٩-١، يتم استخدام المتجه المتوسط لسجلات البيانات في العنقود C_i، باعتباره المركز المتوسط للعنقود لحساب المسافة بين سجل بيانات في العنقود C_i، والمركز المتوسط للعنقود C_i.

حيث إن التعنقد حول K- متوسط يعتمد على المعلمة K فقد تساعد المعرفة بمجال تطبيق الخوارزمية على اختيار قيمة مناسبة لـ K لكي تكون نتائج الخوارزمية ذات معنى ومفيدة في مجال تطبيقها. ويمكن الحصول على نتائج مختلفة من تطبيق الخوارزمية عن طريق استخدام قيم مختلفة لـ K بحيث يمكن مقارنة نتائج تطبيق الخوارزمية.

المثال (٩-١):

استخرج عناقيد حول ٥- متوسطات لمجموعة بيانات اكتشاف أعطال النظام في الجدول ٩-٢ باستخدام المسافة الإقليدية كمقياس للاختلاف. وهذه هي نفس مجموعة البيانات للمثال ٨-١. وتحتوي مجموعة البيانات تسع حالات من الأعطال الآلية الأحادية، وسجل بيانات لكل حالة لها متغيرات الخاصية التسعة عن جودة وحدات المنتج.

في الخطوة ١ من خوارزمية التعنقد حول K- متوسط، نقوم بشكل عشوائي باختيار سجلات البيانات 1، 3، 5، 7 و9 لتجهيز المراكز المتوسطة الأولية للعناقيد الخمسة C_1، C_2، C_3، C_4 وC_5 على التوالي:

الجدول (٩-٢)

مجموعة البيانات لاكتشاف أعطال النظام بتسع حالات من الأعطال الآلية الأحادية

رقم الحالة - Instance (الآلة المعطلة - Faulty Machine)	متغيرات الخاصية عن جودة وحدات المنتج Attribute Variables about Quality of Parts								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1 (M1)	1	0	0	0	1	0	1	0	1
2(M2)	0	1	0	1	0	0	0	1	0
3(M3)	0	0	1	1	0	1	1	1	0
4(M4)	0	0	0	1	0	0	0	1	0
5(M5)	0	0	0	0	1	0	1	0	1
6(M6)	0	0	0	0	0	1	1	0	0
7(M7)	0	0	0	0	0	0	1	0	0
8(M8)	0	0	0	0	0	0	0	1	0
9(M9)	0	0	0	0	0	0	0	0	1

$$\overline{x_{C_1}} = x_1 = \begin{bmatrix}1\\0\\0\\0\\1\\0\\1\\0\\1\end{bmatrix} \quad \overline{x_{C_2}} = x_3 = \begin{bmatrix}0\\0\\1\\1\\0\\1\\1\\1\\0\end{bmatrix} \quad \overline{x_{C_3}} = x_5 = \begin{bmatrix}0\\0\\0\\0\\1\\0\\1\\0\\1\end{bmatrix} \quad \overline{x_{C_4}} = x_7 = \begin{bmatrix}0\\0\\0\\0\\0\\0\\1\\0\\0\end{bmatrix} \quad \overline{x_{C_5}} = x_9 = \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\1\end{bmatrix}.$$

لا تحتوي العناقيد الخمسة على سجلات بيانات في كل منها في البداية. ومن ثم، لدينا $C_1=\{\}$، $C_2=\{\}$، $C_3=\{\}$، $C_4=\{\}$، و $C_5=\{\}$.

في الخطوات ٢ و٣ من الخوارزمية، نأخذ سجل البيانات الأول x_1 من مجموعة البيانات. في الخطوة ٤ من الخوارزمية، نقوم بحساب المسافة الإقليدية لسجل البيانات x_1 إلى كل من العناقيد الخمسة:

$$d\left(x_1, \overline{x_{C_1}}\right)$$

$$= \sqrt{(1-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (1-1)^2}$$

$$= 0$$

$$d\left(x_1, \overline{x_{C_2}}\right)$$

$$= \sqrt{(1-0)^2 + (0-0)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2 + (1-1)^2 + (0-1)^2 + (1-0)^2}$$

$$= 2.65$$

$$d\left(x_1, \overline{x_{C_3}}\right)$$

$$= \sqrt{(1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (1-1)^2}$$

$$= 1$$

$$d\left(x_1, \overline{x_{C_4}}\right)$$

$$= \sqrt{(1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (1-0)^2}$$

$$= 1.73$$

$$d\left(x_1, \overline{x_{C_5}}\right)$$

$= \sqrt{(1-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(1-0)^2+(0-0)^2+(1-0)^2+(0-0)^2+(1-1)^2}$

$= 1.73$

في الخطوة ٥ من الخوارزمية، x_1 لا يتواجد في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية لنقل x_1 إلى أقرب عنقود لها وهو C_1، والذي لا يزال مركزه المتوسط هو نفسه، وذلك لأن مركزه المتوسط تمّ تجهيزه باستخدام x_1 لدينا الآن $C_1 = \{x_1\}$، $C_2 = \{\}$، $C_3 = \{\}$، $C_4 = \{\}$، و$C_5 = \{\}$.

بالعودة إلى الخطوة ٣، نقوم بأخذ سجل البيانات الثاني x_2 من مجموعة البيانات. في الخطوة ٤، نقوم بحساب المسافة الإقليدية لسجل بيانات x_2 إلى كل من العناقيد الخمسة:

$d\left(x_2, \overline{x_{C_1}}\right)$

$= \sqrt{(0-1)^2+(1-0)^2+(0-0)^2+(1-0)^2+(0-1)^2+(0-0)^2+(0-1)^2+(1-0)^2+(0-1)^2}$

$= 2.65$

$d\left(x_2, \overline{x_{C_2}}\right)$

$= \sqrt{(0-0)^2+(1-0)^2+(0-1)^2+(1-1)^2+(0-0)^2+(0-1)^2+(0-1)^2+(1-1)^2+(0-0)^2}$

$= 2$

$d\left(x_2, \overline{x_{C_3}}\right)$

$= \sqrt{(0-0)^2+(1-0)^2+(0-0)^2+(1-0)^2+(0-1)^2+(0-0)^2+(0-1)^2+(1-0)^2+(0-1)^2}$

$= 2.45$

$d\left(x_2, \overline{x_{C_4}}\right)$

$= \sqrt{(0-0)^2+(1-0)^2+(0-0)^2+(1-0)^2+(0-0)^2+(0-0)^2+(0-1)^2+(1-0)^2+(0-0)^2}$

$= 2$

$d\left(x_2, \overline{x_{C_5}}\right)$

$= \sqrt{(0-0)^2+(1-0)^2+(0-0)^2+(1-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(1-0)^2+(0-1)^2}$

$= 2$

في الخطوة ٥، لا يتواجد سجل البيانات x_2 في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية. من بين العناقيد الثلاثة، C_2، C_4 وC_5 والتي تعطي أصغر مسافة لـ x_2 نقوم باختيار C_2 بشكل عشوائي ونقل x_2 إلى C_2. العنقود C_2 يحتوي على سجل بيانات واحد فقط هو x_2 ويتم تحديث المركز المتوسط لـ C_2 من خلال أخذ x_2 كمركزها المتوسط:

$$\overline{x_{C_2}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

لدينا الآن $C_1 = \{x_1\}$، $C_2 = \{x_2\}$، $C_3 = \{\}$، $C_4 = \{\}$، و $C_5 = \{\}$.

بالعودة إلى الخطوة ٣، نأخذ سجل البيانات الثالث x_3 من مجموعة البيانات. في الخطوة ٤، نقوم بحساب المسافة الإقليدية لسجل البيانات x_3 إلى كل من العناقيد الخمسة:

$$d\left(x_3, \overline{x_{C_1}}\right)$$
$$= \sqrt{(0-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (1-0)^2 + (0-1)^2}$$
$$= 2.65$$
$$d\left(x_3, \overline{x_{C_2}}\right)$$
$$= \sqrt{(0-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2}$$
$$= 2$$

$d(x_3, \overline{x_{C_3}})$

$= \sqrt{(0-0)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (1-0)^2 + (0-1)^2}$

$= 2.45$

$d(x_3, \overline{x_{C_4}})$

$= \sqrt{(0-0)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 + (1-1)^2 + (1-0)^2 + (0-0)^2}$

$= 2$

$d(x_3, \overline{x_{C_5}})$

$= \sqrt{(0-0)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2}$

$= 2.45$

في الخطوة ٥، لا يتواجد سجل البيانات x_3 في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية. من بين العنقودين، C_2 وC_4 والتي تعطي أصغر مسافة لـ x_3 نقوم بشكل عشوائي باختيار C_2 ونقل x_3 إلى C_2. العنقود C_2 يحتوي على سجلي بيانات x_2 و x_3 ، ويتم تحديث المركز المتوسط لـ C_2:

$$\overline{x_{C_2}} = \begin{bmatrix} \frac{0+0}{2} \\ \frac{1+0}{2} \\ \frac{0+1}{2} \\ \frac{1+1}{2} \\ \frac{0+0}{2} \\ \frac{0+1}{2} \\ \frac{0+1}{2} \\ \frac{1+1}{2} \\ \frac{0+0}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \\ 1 \\ 0 \\ 0.5 \\ 0.5 \\ 1 \\ 0 \end{bmatrix}$$

لدينا الآن $C_1 = \{ x_1\}$، $C_2 = \{ x_3 x_2\}$، $C_3 = \{ \}$، $C_4 = \{ \}$، و $C_5 = \{ \}$.

بالعودة إلى الخطوة ٣، نأخذ سجل البيانات الرابع x_4 من مجموعة البيانات. في الخطوة ٤، نقوم بحساب المسافة الإقليدية لسجل البيانات x_4 إلى كل من العناقيد الخمسة:

$$d\left(x_4, \overline{x_{C_1}}\right)$$
$$= \sqrt{(0-1)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2}$$
$$= 2.45$$
$$d\left(x_4, \overline{x_{C_2}}\right)$$
$$= \sqrt{(0-0)^2 + (0-0.5)^2 + (0-0.5)^2 + (1-1)^2 + (0-0)^2 + (0-0.5)^2 + (0-0.5)^2 + (1-1)^2 + (0-0)^2}$$
$$= 1$$

$d\left(x_4,\overline{x_{C_3}}\right)$

$= \sqrt{(0-0)^2+(0-0)^2+(0-0)^2+(1-0)^2+(0-1)^2+(0-0)^2+(0-1)^2+(1-0)^2+(0-1)^2}$

$= 2.24$

$d\left(x_4,\overline{x_{C_4}}\right)$

$= \sqrt{(0-0)^2+(0-0)^2+(0-0)^2+(1-0)^2+(0-0)^2+(0-0)^2+(0-1)^2+(1-0)^2+(0-0)^2}$

$= 1.73$

$d\left(x_4,\overline{x_{C_5}}\right)$

$= \sqrt{(0-0)^2+(0-0)^2+(0-0)^2+(1-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(1-0)^2+(0-1)^2}$

$= 1.73$

في الخطوة ٥، لا يتواجد سجل البيانات x_4 في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية لنقل x_4 إلى أقرب عنقود له وهو C_2. ويتم تحديث المركز المتوسط لـ C_2:

$$\overline{x_{C_2}} = \begin{bmatrix} \frac{0+0+0}{3} \\ \frac{1+0+0}{3} \\ \frac{0+1+0}{3} \\ \frac{1+1+1}{3} \\ \frac{0+0+0}{3} \\ \frac{0+1+0}{3} \\ \frac{0+1+0}{3} \\ \frac{1+1+1}{3} \\ \frac{0+0+0}{3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.33 \\ 0.33 \\ 1 \\ 0 \\ 0.33 \\ 0.33 \\ 1 \\ 0 \end{bmatrix}.$$

لدينا الآن $C_1 = \{x_1\}$، $C_2 = \{x_2, x_3, x_4\}$، $C_3 = \{\ \}$، $C_4 = \{\ \}$، و $C_5 = \{\ \}$.

بالعودة إلى الخطوة ٣، نأخذ سجل البيانات الخامس x_5 من مجموعة البيانات. في الخطوة ٤، نعلم أن x_5 هو الأقرب إلى C_3 حيث أنه تمّ تشكيل C_3 في البداية باستخدام C_5 ولم يتم تحديثه منذ ذلك الحين. في الخطوة ٥، لا يتواجد x_5 في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية لنقل x_5 إلي أقرب عنقود له وهو C_3 والذي لا يزال مركزه المتوسط هو نفسه.

لدينا الآن $C_1 = \{x_1\}$، $C_2 = \{x_2, x_3, x_4\}$، $C_3 = \{x_5\}$، $C_4 = \{\ \}$، $C_5 = \{\ \}$.

بالعودة إلى الخطوة ٣، نأخذ سجل البيانات السادس x_6 من مجموعة البيانات. في الخطوة ٤، نقوم بحساب المسافة الإقليدية لسجل البيانات x_6 إلى كل من المجموعات الخمسة:

$$d\left(x_6, \overline{x_{C_1}}\right)$$

$$= \sqrt{(0-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2 + (0-1)^2}$$

$$= 2$$

$$d\left(x_6, \overline{x_{C_2}}\right)$$

$$= \sqrt{(0-0)^2 + (0-0.33)^2 + (0-0.33)^2 + (0-1)^2 + (0-0)^2 + (1-0.33)^2 + (1-0.33)^2 + (0-1)^2 + (0-0)^2}$$

$$= 1.77$$

$$d\left(x_6, \overline{x_{C_3}}\right)$$

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2 + (0-1)^2}$$

$$= 1.73$$

$$d\left(x_6, \overline{x_{C_4}}\right)$$

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2}$$

$$= 1$$

$$d\left(x_6, \overline{x_{C_5}}\right)$$

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2}$$

$$= 1.73$$

في الخطوة ٥، لا يتواجد سجل البيانات x_6 في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية لنقل x_6 إلى أقرب عنقود له وهو C_4 ويتم تحديث المركز المتوسط لـ C_4:

$$\overline{x_{C_4}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

لدينا الآن $C_1 = \{x_1\}$، $C_2 = \{x_2, x_3, x_4\}$، $C_3 = \{x_5\}$، $C_4 = \{x_6\}$، $C_5 = \{\}$.

بالعودة إلى الخطوة ٣، نأخذ سجل البيانات السابع x_7 من مجموعة البيانات. في الخطوة ٤، نقوم بحساب المسافة الإقليدية لسجل البيانات x_7 إلى كل من العناقيد الخمسة:

$$d\left(x_7, \overline{x_{C_1}}\right)$$

$$= \sqrt{(0-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (0-1)^2}$$

$$= 1.73$$

$$d\left(x_7, \overline{x_{C_2}}\right)$$

$$= \sqrt{(0-0)^2 + (0-0.33)^2 + (0-0.33)^2 + (0-1)^2 + (0-0)^2 + (0-0.33)^2 + (1-0.33)^2 + (0-1)^2 + (0-0)^2}$$

$$= 1.67$$

$$d\left(x_7, \overline{x_{C_3}}\right)$$

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (0-1)^2}$$

$$= 1.41$$

$$d\left(x_7, \overline{x_{C_4}}\right)$$

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2}$$

$$= 1$$

$$d\left(x_7, \overline{x_{C_5}}\right)$$

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2}$$

$$= 1.41$$

في الخطوة ٥، لا يتواجد سجل البيانات x_7 في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية لنقل x_7 إلى أقرب عنقود له وهو C_4 ويتم تحديث المركز المتوسط لـ C_4:

$$\overline{x_{C_4}} = \begin{bmatrix} \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \\ \frac{1+0}{2} \\ \frac{1+1}{2} \\ \frac{0+0}{2} \\ \frac{0+0}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

لدينا الآن $C_1 = \{x_1\}$، $C_2 = \{x_2, x_3, x_4\}$، $C_3 = \{x_5\}$، $C_4 = \{x_6, x_7\}$، $C_5 = \{\ \}$.

بالعودة إلى الخطوة ٣، نأخذ سجل البيانات الثامن x_8 من مجموعة البيانات. في الخطوة ٤، نقوم بحساب المسافة الإقليدية لسجل البيانات x_8 إلى كل من العناقيد الخمسة:

$$d\left(x_8, \overline{x_{C_1}}\right)$$
$$= \sqrt{(0-1)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-1)^2+(0-0)^2+(0-1)^2+(1-0)^2+(0-1)^2}$$
$$= 2.27$$

$$d\left(x_8, \overline{x_{C_2}}\right)$$
$$= \sqrt{(0-0)^2+(0-0.33)^2+(0-0.33)^2+(0-1)^2+(0-0)^2+(0-0.33)^2+(0-0.33)^2+(1-1)^2+(0-0)^2}$$
$$= 1.20$$

$$d\left(x_8, \overline{x_{C_3}}\right)$$
$$= \sqrt{(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-1)^2+(0-0)^2+(0-1)^2+(1-0)^2+(0-1)^2}$$
$$= 2$$

$$d\left(x_8, \overline{x_{C_4}}\right)$$
$$= \sqrt{(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0.5)^2+(0-1)^2+(1-0)^2+(0-0)^2}$$
$$= 1.5$$

$$d\left(x_8, \overline{x_{C_5}}\right)$$
$$= \sqrt{(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(1-0)^2+(0-1)^2}$$
$$= 1.41$$

في الخطوة ٥، لا يتواجد سجل البيانات x_8 في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية لنقل x_8 إلى أقرب عنقود له وهو C_2 ويتم تحديث المركز المتوسط لـ C_2:

$$\overline{x_{C_2}} = \begin{bmatrix} \frac{0+0+0+0}{4} \\ \frac{1+0+0+0}{4} \\ \frac{0+1+0+0}{4} \\ \frac{1+1+1+0}{4} \\ \frac{0+0+0+0}{4} \\ \frac{0+1+0+0}{4} \\ \frac{0+1+0+0}{4} \\ \frac{1+1+1+1}{4} \\ \frac{0+0+0+0}{4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.25 \\ 0.25 \\ 0.75 \\ 0 \\ 0.25 \\ 0.25 \\ 1 \\ 0 \end{bmatrix}.$$

لدينا الآن $C_1 = \{x_1\}$، $C_2 = \{x_2 x_3, x_4, x_8\}$ ، $C_3 = \{x_5\}$، $C_4 = \{x_6, x_7\}$ ، $C_5 = \{\ \}$.

بالعودة إلى الخطوة ٣، نأخذ سجل البيانات التاسع x_9 من مجموعة البيانات. في الخطوة ٤، نعلم أن x_9 هو الأقرب إلى C_5. لأنه تمّ إنشاء C_5 باستخدام x_9 ولم يتم تحديثه منذ ذلك الحين. في الخطوة ٥، لا يتواجد x_9 في أي عنقود. يتم تنفيذ الخطوة ٦ من الخوارزمية لنقل x_9 إلى أقرب عنقود له والذي لا يزال مركزه المتوسط هو نفسه.

لدينا الآن $C_1 = \{x_1\}$، $C_2 = \{x_2, x_3, x_4, x_8\}$، $C_3 = \{x_5\}$، $C_4 = \{x_6, x_7\}$، $C_5 = \{x_9\}$.

بعد الانتهاء من تنفيذ تعليمة *(FOR)* في الخطوات ٣-٦، نذهب إلى الخطوة ٧. نظراً لأن هناك تغييرات على المركز المتوسط للعنقود في الخطوات ٣-٦، نعود إلى الخطوة ٢ ثم الخطوة ٣ لبدء تكرار آخر لتعليمة (*FOR*). في تعليمة (*FOR*) هذه، يكون العنقود الحالي لكل سجل بيانات هو العنقود الأقرب لسجل البيانات. ومن ثم، فإنه لا ينتقل سجل من سجلات البيانات التسعة من عنقوده الحالي إلى عنقود آخر، ولا يحدث أي تغيير للمركز المتوسط للعنقود في تعليمة (*FOR*) هذه. إن العناقيد حول ٥ متوسطات في هذا المثال ينتج عنها ٥ عناقيد، $C_1 = \{x_1\}$، $C_2 = \{x_2, x_3, x_4, x_8\}$، $C_3 = \{x_5\}$، $C_4 = \{x_6, x_7\}$، و $C_5 = \{x_9\}$. وينتج التعنقُد الهرمي لنفس مجموعة البيانات في الشكل ٨-١ خمس عناقيد، $\{x_1, x_5\}$، $\{x_2, x_4, x_8\}$، $\{x_3\}$، $\{x_6, x_7\}$، و$\{x_9\}$، عندما وضعنا حداً لمسافة الدمج تساوي القيمة 1.5. من ثم، فإن نتائج التعنقد حول ٥ متوسطات متشابهة ولكنها ليست بالضبط نتائج التعنقد الهرمي نفسه.

٩-٢ التعنقد القائم على الكثافة (Density-Based Clustering):

يعد التعنقد القائم على الكثافة أن عناقيد البيانات عبارة عن مناطق سجلات البيانات بكثافة عالية، والتي يتم قياسها باستخدام عدد سجلات البيانات داخل نصف قطر محدد (*Li and Ye, 2002*). يتم فصل العناقيد حسب مناطق سجلات البيانات المنخفضة الكثافة. الخوارزمية *DBSCAN* (*Ester et al., 1996*) عبارة عن خوارزمية التعنقد القائم على الكثافة التي تبدأ بمجموعة من سجلات البيانات ومعلمتين (*two parameters*) هما: نصف القطر والحد الأدنى من عدد سجلات البيانات المطلوب لتشكيل عنقود واحد. يتم حساب كثافة سجل البيانات x عن طريق حساب عدد سجلات البيانات داخل نصف قطر سجل البيانات x. إنَّ منطقة x تمثل المساحة داخل نصف قطر x والتي يتم اعتبار أن لها منطقة كثيفة إذا كان عدد سجلات البيانات في المنطقة x أكبر أو يساوي الحد الأدنى من عدد سجلات البيانات. في البداية، يتم اعتبار جميع سجلات البيانات في مجموعة البيانات غير مُعلَّمة. تختار خوارزمية التعنقد القائم على الكثافة (*DBSCAN*) بصورة عشوائية سجل بيانات غير مُعلم x من مجموعة البيانات. إذا كانت منطقة سجل البيانات x غير كثيفة، يتم وضع علامة على سجل البيانات x باعتباره سجل ضوضاء (*noise data point*). إذا كانت منطقة x كثيفة، يتم تشكيل عنقود جديد يحتوي على x ويتم وضع علامة على x باعتباره

عضواً في هذا العنقود الجديد. علاوةً على ذلك، ينضم كل من سجلات البيانات الأخرى والموجودة في منطقة x إلى العنقود ويتم وضع علامة عليه بوصفه عضواً في هذا العنقود إذا لم يكن سجل البيانات هذا قد انضم بعد إلى أي عنقود. يتم توسيع هذا العنقود الجديد ليشمل جميع سجلات البيانات التي لم تنضم بعد إلى عنقود معين والتي تكون في المنطقة الخاصة بسجل بيانات معين، وليكن z، والذي هو موجود في العنقود إذا كانت منطقة z كثيفة. ويستمر التوسع في العنقود حتى تنضم جميع سجلات البيانات المتصلة من خلال المناطق الكثيفة لسجلات البيانات إلى العنقود إذا لم تكن قد انضمت بعد إلى العنقود. نلاحظ أن سجل بيانات الضوضاء قد يكون موجوداً في وقت لاحق في المنطقة الكثيفة لسجل بيانات معين في عنقود آخر، ومن ثم يمكن تحويله إلى عضو في ذلك العنقود. بعد اكتمال العنقود، تختار خوارزمية التعنقد القائم على الكثافة (*DBSCAN*) سجل بيانات آخر غير مُعلّم وتُقيّم الخوارزمية ما إذا كان سجل بيانات هذا عبارة عن سجل ضوضاء أو سجل بيانات يتم البدء به لبناء عنقود جديد. وتستمر هذه العملية حتى يتم تعليم كافة سجلات البيانات في مجموعة البيانات إما كسجل ضوضاء أو كعضو في عنقود.

بما أن التعنقد القائم على الكثافة يعتمد على معلمتين هما نصف القطر والحد الأدنى لعدد سجلات البيانات، فإن المعرفة بمجال التطبيق المبحوث والمستهدف قد يساعد على اختيار قيم مناسبة للمعلمتين للحصول على نتيجة تعنقد ذات معنى في مجال التطبيق. ويمكن الحصول على نتائج تعنقد مختلفة باستخدام قيم معلمات مختلفة بحيث يمكن مقارنة النتائج المختلفة وتقييمها.

٩-٣ البرمجيات والتطبيقات (Software and Applications):

تمّ دعم استخدام التعنقد حول K- متوسط في كل من البرمجيات التالية:

- *WEKA (http://www.cs.waikato.ac.nz/ml/weka/*)
- *MATLAB* (*www.matworks.com*).
- *SAS* (*www.sas.com*).

يمكن الحصول على تطبيق واستخدام خوارزمية التعنقد القائم على الكثافة (*DBSCAN*) للبيانات المكانية (*spatial data*) في (*Ester et al., 1996*).

التمارين (Exercises):

٩-١ استخرج تعنقداً حول متوسطين *(2-means)* من سجلات البيانات في الجدول ٩-٢ باستخدام المسافة الإقليدية كمقياس للاختلاف وباستخدام سجلات البيانات الأولى والثالثة لتجهيز المراكز المتوسطة الأولية للعنقودين.

٩-٢ استخرج التعنقد القائم على الكثافة لسجلات البيانات في الجدول ٩-٢ باستخدام المسافة الإقليدية كمقياس للاختلاف، ويكون *1.5* هو نصف القطر و*2* هو الحد الأدنى لعدد سجلات البيانات المطلوبة لتشكيل عنقود معين.

٩-٣ استخرج التعنقد القائم على الكثافة لسجلات البيانات في الجدول ٩-٢ باستخدام المسافة الإقليدية كمقياس للاختلاف، ويكون *2* هو نصف القطر و*2* هو الحد الأدنى لعدد سجلات البيانات المطلوبة لتشكيل عنقود معين.

٩-٤ استخرج تعنقداً حول ٣- متوسطات لـ ٢٣ سجل من سجلات البيانات في مجموعة البيانات الدائرية في مكوك الفضاء الواردة في الجدول ١-٢. قم باستخدام درجة حرارة الإطلاق (*Launch- Temperatue*) وضغط التحقق من التسرب (*Leak-Check Pressure*) باعتبارها متغيرات الخاصية ودالة التطبيع في المعادلة ٧-٤ للحصول على قيم مطبّعة لدرجة حرارة الإطلاق وضغط التحقق من التسرب أيضاً، والمسافة الإقليدية كمقياس للاختلاف.

٩-٥ كرّر التمرين ٩-٤ باستخدام مقياس تشابه جيب التمام (جتا) (*cosine similarity measure*).

١٠- خريطة التنظيم الذاتي

Self-Organizing Map - SOM

يستعرض هذا الفصل خريطة التنظيم الذاتي (*Self – Organizing MAP – SOM*)، والتي تقوم على أساس المعمارية الخاصة بالشبكات العصبية الصناعية (*ANN*)، وتُستخدَم خريطة التنظيم الذاتي لغرض عنقدة وتصوير البيانات. تمّ سرد قائمة من حزم البرمجيات الخاصة بخريطة التنظيم الذاتي (*SOM*) إلى جانب المراجع للتطبيقات.

١٠-١ خوارزمية خريطة التنظيم الذاتي (Algorithm of Self-Organizing Map):

طوّر كونين (*Kohonen*) سنة ١٩٨٢م خريطة التنظيم الذاتي (*SOM*). وهي عبارة عن شبكة عصبية صناعية (*ANN*) بعقد مخرجات (*output nodes*) مرتبة ومنظمة في فضاء يحتوي على q – من الأبعاد، وتُسمى هذه الشبكة بخريطة المخرجات (*output map*)، أو الرسم البياني (*graph*). وعادةً ما يُستخدَم فضاء أحادي أو ثنائي أو ثلاثي الأبعاد، أو ترتيب مُعيّن لعُقَد المخرجات، كما هو مبين في الشكل ١٠-١، ومن ثم يكون من الممكن تصور وتخيل عناقيد سجلات البيانات، لأنه يتم تمثيل السجلات المتشابهة على شكل عُقَد (*nodes*) قريبة من بعضها البعض في خريطة المخرجات.

في أي خريطة تنظيم ذاتي (*SOM*)، يتم ربط كل متغير من متغيرات المدخلات، x_i، $i= 1,\dots,p$، بكل عقدة في خريطة التنظيم الذاتي (*SOM*) j، $j= 1,\dots,k$ مع وزن لهذا الارتباط يرمز له بـ w_{ji}. يتم حساب متجه المخرجات (*output vector*)، ويرمز له بـ o، الخاص بخريطة التنظيم الذاتي (*SOM*) لمتجه مدخلات معطى x على النحو التالي:

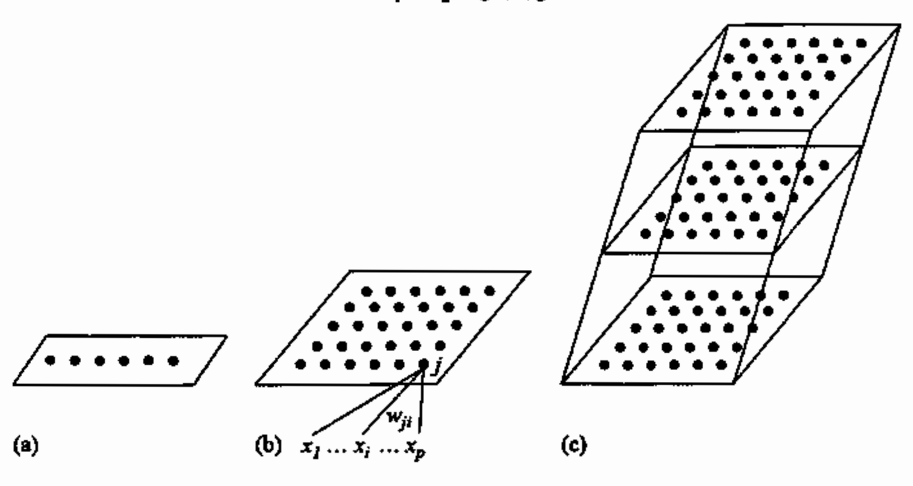
$$o = \begin{bmatrix} o_1 \\ \vdots \\ o_j \\ \vdots \\ o_j \end{bmatrix} = \begin{bmatrix} w_1'x \\ \vdots \\ w_j'x \\ \vdots \\ w_k'x \end{bmatrix}, \qquad (١٠-١)$$

حيث إن:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_p \end{bmatrix}$$

الشكل (١٠-١)

التصاميم الخاصة بخريطة التنظيم الذاتي (SOM) بخريطة مخرجات (a) أحادية، (b) ثنائية، و(c) وثلاثية الأبعاد



$$w_j = \begin{bmatrix} w_{j1} \\ \vdots \\ w_{ji} \\ \vdots \\ w_{jp} \end{bmatrix}.$$

من بين جميع عُقَد المخرجات، تُسمى عقدة المخرجات التي تعطي أكبر قيمة لمتجه مدخلات معطى x بالعقدة الفائزة (*winner node*). يكون للعقدة الفائزة الخاصة بمتجه

المدخلات متجه وزن أكثر مماثلةً ومشابهةً لمتجه المدخلات. تحدد خوارزمية التعلم لخريطة التنظيم الذاتي (SOM) أوزان الارتباط بحيث تكون العُقَد الفائزة لمتجهات المدخلات المتشابهة قريبةً بعضها من بعض. يوضح الجدول ١٠-١ خطوات خوارزمية التعلم لخريطة التنظيم الذاتي (SOM) إذا كان لدينا مجموعة بيانات تدريبية أو استكشافية بعدد n من نقاط البيانات، x_i $i=1,\ldots,n$.

في الخطوة ٥ من الخوارزمية، يتم تحديث أوزان الارتباط للعقدة الفائزة لمتجه المدخلات x_i والعُقَد المجاورة من العقدة الفائزة لجعل أوزان العقدة الفائزة والعُقَد المجاورة لها أكثر مماثلةً ومشابهةً لمتجه المدخلات، ومن ثم جعل هذه العقد تقوم بإنتاج مخرجات أكبر لمتجه المدخلات. يمكن تعريف دالة الجوار $f(j,c)$ والتي تحدد مدى قرب العقدة j إلى العقدة الفائزة c ومن ثم أهلية العقدة j لتغيير الوزن، بطرق عديدة. أحد الأمثلة على دالة الجوار $f(j,c)$:

$$f(j,c)=\begin{cases}1 & \text{if}\left\|r_j-r_c\right\|\leq B_c(t)\\ 0 & \text{otherwise}\end{cases}, \qquad (١٠\text{-}٢)$$

حيث r_j وr_c هي إحداثيات العقدة j والعقدة الفائزة c في خريطة المخرجات، وتمثل $B_c(t)$ قيمة الحد التي تقيد مدى الجوار من العقدة الفائزة c.

الجدول (١٠-١) خوارزمية التعلم لخريطة التنظيم الذاتي (SOM) - (إنجليزي وعربي)

Step	Description
1	Initialize the connection weights of nodes with random positive or negative values, $w'_j(t)=\left[w_{j1}(t) \quad \cdots \quad w_{jp}(t)\right], t=0, j=1, \ldots, k$
2	REPEAT
3	FOR $i = 1$ to n
4	Determine the winner node c for x_i: $c = \mathrm{argmax}_j w'_j(t)x_i$
5	Update the connection weights of the winner node and its nearby nodes: $w_j(t+1)=w_j(t)+\alpha f(j,c)\left[x_i - w_j(t)\right]$, where α is the learning rate and $f(j, c)$ defines whether or not node j is close enough to c to be considered for the weight update
6	$w_j(t+1)=w_j(t)$ for other nodes without the weight update
7	$t = t + 1$
8	UNTIL the sum of weight changes for all the nodes, $E(t)$, is not greater than a threshold ε

الخطوة	الوصف
١	جهز قيماً أولية لأوزان الارتباط للعُقَد بقيم عشوائية موجبة أو سالبة $w'_j(t)= [w_{j1}(t) \ldots w_{jp}(t)], t=0, j=1, \ldots, k$
٢	كرر (REPEAT)
٣	كرر (FOR) إبتداء من $i=1$ إلى n
٤	حدد العقدة الفائزة c لـ x_i: $c=\mathrm{argmax}_j\, w'_j(t)\, x_i$
٥	حدّث أوزان الارتباط للعقدة الفائزة والعُقَد المجاورة لها: $w_j(t+1)= w_j(t) + \alpha f(j,c)[x_i - w_j(t)]$ حيث إن α هي معدل التعلّم و $f(j,c)$ تعرف ما إذا كانت العقدة j قريبةً بما فيه الكفاية إلى c حتى يتم أخذها في الاعتبار أثناء تحديث الأوزان.
٦	اجعل $w_j(t+1) = w_j(t)$ للعقد الأخرى دون تحديث الوزن
٧	$t = t+1$
٨	شرط توقف التكرار (UNTIL) : لا يكون مجموع تغيرات الوزن لكل العُقَد، $E(t)$، أكبر من الحد ε

يتم تعريف $B_c(t)$ كدالة لـ t بحيث تستخدم عملية تعلُّم تكيفي والتي تستخدم قيمة حد كبيرة في بداية عملية التعلُّم، ومن ثمَّ يتم تخفيض قيم الحد مع كل تكرار في الخوارزمية.
مثال آخر للدالة $f(j,c)$ هو:

$$f(j,c) = \frac{1}{\frac{\left\|r_j - r_c\right\|^2}{e^{2B_c^2(t)}}}. \qquad (٣-١٠)$$

في الخطوة ٨ من الخوارزمية، يتم حساب مجموع تغييرات الوزن لكافة العُقَد:

$$E(t) = \sum_j \left\|w_j(t+1) - w_j(t)\right\|. \qquad (٤-١٠)$$

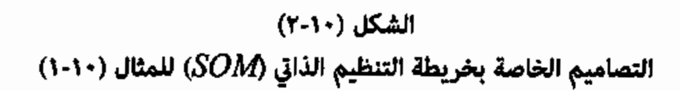
بعد أن يتم تعلُّم خريطة التنظيم الذاتي (*SOM*)، يتم تحديد عناقيد سجلات البيانات عن طريق وضع علامة على كل عقدة ذات سجل البيانات (أو سجلات البيانات) التي تجعل تلك العقدة هي العقدة الفائزة. ويتم معرفة وتحديد موقع عنقود سجلات البيانات بحيث يكون في منطقة مجاورة وقريبة في خريطة المخرجات.

المثال (١٠-١):

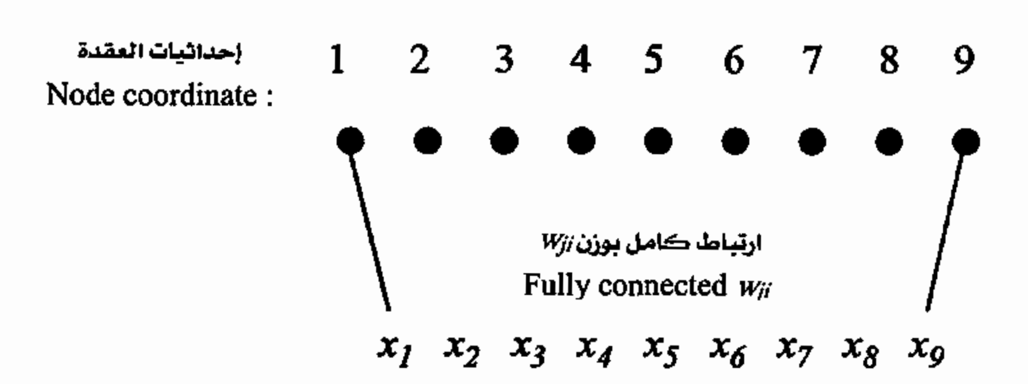
استخدام خريطة التنظيم الذاتي (*SOM*) بتسع عقد في سلسلة أحادية الأبعاد، وتكون إحداثيات العقد كالتالي: 1، 2، 3، 4، 5، 6، 7، 8، و9، كما في الشكل ١٠-٢، لتجميع وعنقدة نقاط البيانات التسعة الموجودة في مجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع في الجدول ١٠-٢، وهي نفس مجموعة البيانات في الجداول ٨-١ و٩-٢. وتحتوي مجموعة البيانات على تسع حالات للأعطال الآلية الأحادية، ويحتوي سجل البيانات لكل حالة على تسعة متغيرات خاصية خاصة بجودة وحدات المنتج. معدل التعلم α هو 0.3.
ودالة الجوار $f(j, c)$ هي:

$$f(j,c) = \begin{cases} 1 & \text{for } j = c-1, c, c+1 \\ 0 & \text{otherwise} \end{cases},$$

الشكل (١٠-٢)



التصاميم الخاصة بخريطة التنظيم الذاتي (*SOM*) للمثال (١٠-١)



الجدول (١٠-٢)

مجموعة البيانات الخاصة بالكشف عن أعطال نظام التصنيع بتسع حالات للأعطال الآلية الأحادية

متغيرات الخاصية عن جودة وحدات المنتج Attribute Variables about Quality of Parts									رقم الحالة - Instance (الآلة المعطلة - Faulty Machine)
x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	
1	0	1	0	1	0	0	0	1	1 (M1)
0	1	0	0	0	1	0	1	0	2(M2)
0	1	1	1	0	1	1	0	0	3(M3)
0	1	0	0	0	1	0	0	0	4(M4)
1	0	1	0	1	0	0	0	0	5(M5)
0	0	1	1	0	0	0	0	0	6(M6)
0	0	1	0	0	0	0	0	0	7(M7)
0	1	0	0	0	0	0	0	0	8(M8)
1	0	0	0	0	0	0	0	0	9(M9)

في الخطوة ١ من عملية التعلُّم، نقوم بتهيئة أوزان الارتباط بالقيم الأولية العشوائية التالية:

$$w_1(0) = \begin{bmatrix} -0.24 \\ -0.41 \\ 0.46 \\ 0.27 \\ 0.88 \\ -0.09 \\ 0.78 \\ -0.39 \\ 0.91 \end{bmatrix} \quad w_2(0) = \begin{bmatrix} 0.44 \\ 0.44 \\ 0.93 \\ -0.15 \\ 0.84 \\ -0.36 \\ -0.16 \\ 0.55 \\ 0.93 \end{bmatrix} \quad w_3(0) = \begin{bmatrix} 0.96 \\ -0.45 \\ -0.75 \\ 0.35 \\ 0.05 \\ 0.86 \\ 0.12 \\ -0.49 \\ 0.98 \end{bmatrix} \quad w_4(0) = \begin{bmatrix} 0.82 \\ -0.22 \\ 0.60 \\ -0.56 \\ 0.91 \\ -0.80 \\ 0.33 \\ -0.54 \\ 0.47 \end{bmatrix}$$

$$w_5(0) = \begin{bmatrix} 0.62 \\ 0.44 \\ 0.33 \\ 0.46 \\ -0.25 \\ -0.26 \\ -0.71 \\ -0.61 \\ 0.38 \end{bmatrix} \quad w_6(0) = \begin{bmatrix} -0.47 \\ -0.62 \\ -0.96 \\ -0.43 \\ 0.32 \\ 0.96 \\ 0.70 \\ -0.04 \\ -0.84 \end{bmatrix} \quad w_7(0) = \begin{bmatrix} -0.87 \\ 0.23 \\ 0.37 \\ 0.49 \\ 0.04 \\ 0.33 \\ -0.10 \\ 0.45 \\ -0.96 \end{bmatrix}$$

$$w_8(0) = \begin{bmatrix} -0.95 \\ -0.21 \\ -0.48 \\ 0.05 \\ -0.54 \\ 0.23 \\ -0.37 \\ 0.61 \\ -0.76 \end{bmatrix} \quad w_9(0) = \begin{bmatrix} 0.69 \\ 0.23 \\ -0.69 \\ 0.86 \\ 0.22 \\ -0.91 \\ 0.82 \\ 0.31 \\ 0.31 \end{bmatrix}$$

استخدام هذه الأوزان الأولية لحساب مخرجات خريطة التنظيم الذاتي (SOM) لسجلات البيانات التسعة يجعل العقد أرقام 4، 9، 7، 9، 1، 6، 9، 8، و3، هي العُقَد الفائزة لـ x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 وx_9 على التوالي. على سبيل المثال، يتم حساب المخرجات الخاصة بكل عقدة x_1 لتحديد العقدة الفائزة:

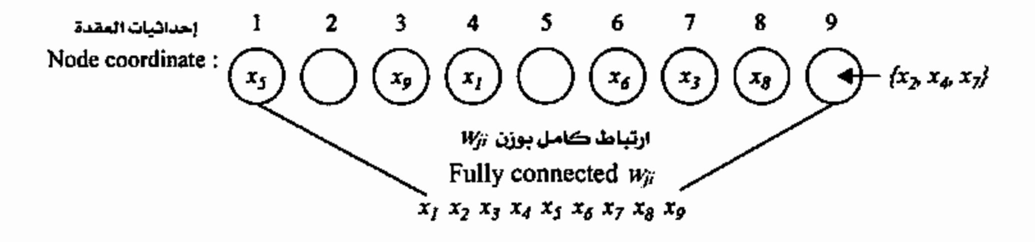
$$o = \begin{bmatrix} o_1 \\ o_2 \\ o_3 \\ o_4 \\ o_5 \\ o_6 \\ o_7 \\ o_8 \\ o_9 \end{bmatrix} = \begin{bmatrix} w_1'(0)x_1 \\ w_2'(0)x_1 \\ w_3'(0)x_1 \\ w_4'(0)x_1 \\ w_5'(0)x_1 \\ w_6'(0)x_1 \\ w_7'(0)x_1 \\ w_8'(0)x_1 \\ w_9'(0)x_1 \end{bmatrix}$$

$$= \begin{bmatrix} (-0.24)(1) + (-0.41)(0) + (0.46)(0) + (0.27)(0) + (0.88)(1) + (-0.09)(0) \\ +(0.78)(1) + (-0.39)(0) + (0.91)(1) \\ \\ (0.44)(1) + (0.44)(0) + (0.93)(0) + (-0.15)(0) + (0.84)(1) + (-0.36)(0) \\ +(-0.16)(1) + (0.55)(0) + (0.93)(1) \\ \\ (0.96)(1) + (-0.45)(0) + (-0.75)(0) + (0.75)(0) + (0.05)(1) + (0.86)(0) \\ +(0.12)(1) + (-0.49)(0) + (0.98)(1) \\ \\ (0.82)(1) + (-0.22)(0) + (0.60)(0) + (-0.56)(0) + (0.91)(1) + (-0.89)(0) \\ +(0.33)(1) + (-0.54)(0) + (0.47)(1) \\ \\ (0.62)(1) + (0.44)(0) + (0.33)(0) + (0.46)(0) + (-0.25)(1) + (-0.26)(0) \\ +(-0.71)(1) + (-0.61)(0) + (0.38)(1) \\ \\ (-0.47)(1) + (-0.62)(0) + (-0.96)(0) + (-0.43)(0) + (0.32)(1) + (0.96)(0) \\ +(0.70)(1) + (-0.04)(0) + (-0.84)(1) \\ \\ (-0.87)(1) + (0.23)(0) + (0.37)(0) + (0.49)(0) + (0.04)(1) + (0.33)(0) \\ +(-0.10)(1) + (0.45)(0) + (-0.96)(1) \\ \\ (-0.95)(1) + (-0.21)(0) + (-0.48)(0) + (0.05)(0) + (-0.54)(1) + (0.23)(0) \\ +(-0.37)(1) + (0.61)(0) + (-0.76)(1) \\ \\ (0.69)(1) + (0.23)(0) + (-0.69)(0) + (0.86)(0) + (0.22)(1) + (-0.91)(0) \\ +(0.82)(1) + (0.31)(0) + (0.31)(1) \end{bmatrix}$$

$$= \begin{bmatrix} 2.33 \\ 2.04 \\ 2.11 \\ 2.53 \\ 0.04 \\ -0.29 \\ -1.90 \\ -2.62 \\ 2.04 \end{bmatrix}$$

الشكل (١٠-٣)

العقد الفائزة لنقاط البيانات التسع في المثال (١٠-١) باستخدام قيم الوزن أولية



وحيث إن العقدة رقم 4 لها أكبر قيمة مخرجات $o_4 = 2.53$، فإن العقدة 4 هي العقدة الفائزة لـ x_1. يوضح الشكل ١٠-٣ خريطة المخرجات للإشارة إلى العقدة الفائزة لسجلات البيانات التسع، ومن ثم يكون لدينا عناقيد أولية لسجلات البيانات على أساس الأوزان الأولية.

في الخطوات ٢ و٣، يؤخذ في الاعتبار سجل البيانات x_1 في الخطوة ٤، يتم حساب المخرجات الخاصة بكل عقدة لـ x_1 لتحديد العقدة الفائزة. كما هو موضح سابقاً، فإن العقدة 4 هي العقدة الفائزة لـ x_1 ومن ثم، $c=4$. وفي الخطوة ٥، يتم تحديث أوزان الارتباط إلى العقدة الفائزة $c=4$ ومجاوراتها $c-1=3$ و $c+1=5$:

$$w_4(1) = w_4(0) + (0.3)[x_1 - w_4(0)] = (0.7)w_4(0) + (0.3)x_1$$

$$= (0.7)\begin{bmatrix} 0.82 \\ -0.22 \\ 0.60 \\ -0.56 \\ 0.91 \\ -0.80 \\ 0.33 \\ -0.54 \\ 0.47 \end{bmatrix} + (0.3)\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.87 \\ -0.15 \\ 0.42 \\ -0.39 \\ 0.94 \\ -0.56 \\ 0.53 \\ -0.38 \\ 0.63 \end{bmatrix}.$$

$$w_3(1) = w_3(0) + (0.3)[x_1 - w_3(0)] = (0.7)w_3(0) + (0.3)x_1$$

$$= (0.7)\begin{bmatrix} 0.96 \\ -0.45 \\ -0.75 \\ 0.35 \\ 0.05 \\ 0.86 \\ 0.12 \\ -0.49 \\ 0.98 \end{bmatrix} + (0.3)\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.96 \\ -0.32 \\ 0.53 \\ 0.25 \\ 0.34 \\ 0.60 \\ 0.38 \\ -0.34 \\ 0.99 \end{bmatrix}.$$

$$w_5(1) = w_5(0) + (0.3)[x_1 - w_5(0)] = (0.7)w_5(0) + (0.3)x_1$$

$$= (0.7)\begin{bmatrix}0.62\\0.44\\0.33\\0.46\\-0.25\\-0.26\\-0.71\\-0.61\\0.38\end{bmatrix} + (0.3)\begin{bmatrix}1\\0\\0\\0\\1\\0\\1\\0\\1\end{bmatrix} = \begin{bmatrix}0.73\\0.31\\0.23\\0.32\\0.13\\-0.18\\0.80\\-0.43\\0.57\end{bmatrix}.$$

في الخطوة ٦، تبقى الأوزان الخاصة بالعُقَد الأخرى هي نفسها. وفي الخطوة ٧، تزيد t إلى 1، وتكون الأوزان الخاصة بالعقد التسعة هي:

$$w_1(1) = \begin{bmatrix}-0.24\\-0.41\\0.46\\0.27\\0.88\\-0.09\\0.78\\-0.39\\0.91\end{bmatrix} \quad w_2(1) = \begin{bmatrix}0.44\\0.44\\0.93\\-0.15\\0.84\\-0.36\\-0.16\\0.55\\0.93\end{bmatrix} \quad w_3(1) = \begin{bmatrix}1.96\\-0.32\\0.53\\0.25\\0.34\\0.60\\0.38\\-0.34\\0.99\end{bmatrix} \quad w_4(1) = \begin{bmatrix}0.87\\-0.15\\0.42\\-0.39\\0.94\\-0.56\\0.53\\-0.38\\0.63\end{bmatrix}$$

$$w_5(1) = \begin{bmatrix}0.73\\0.31\\0.23\\0.32\\0.13\\-0.18\\0.80\\-0.43\\0.57\end{bmatrix} \quad w_6(1) = \begin{bmatrix}-0.47\\-0.62\\-0.96\\-0.43\\0.32\\0.96\\0.70\\-0.04\\-0.84\end{bmatrix} \quad w_7(1) = \begin{bmatrix}-0.87\\0.23\\0.37\\0.49\\0.04\\0.33\\-0.10\\0.45\\-0.96\end{bmatrix}$$

$$w_8(1) = \begin{bmatrix} -0.95 \\ -0.21 \\ -0.48 \\ 0.05 \\ -0.54 \\ 0.23 \\ -0.37 \\ 0.61 \\ -0.76 \end{bmatrix} \quad w_9(1) = \begin{bmatrix} 0.69 \\ 0.23 \\ -0.69 \\ 0.86 \\ 0.22 \\ -0.91 \\ 0.82 \\ 0.31 \\ 0.31 \end{bmatrix}$$

بعد ذلك، نعود إلى الخطوات ٢ و٣، ويؤخذ في الاعتبار سجل البيانات x_2 وتتواصل عملية التعلم حتى يصبح مجموع التغييرات المتعاقبة للأوزان، والتي استهلتها كل سجلات البيانات التسع، صغيرة بما فيه الكفاية.

٢-١٠ البرامج والتطبيقات (Software and Applications):

يتم دعم خريطة التنظيم الذاتي (*SOM*) عن طريق البرمجيات:

- *Weka (http://www.cs.waikato.ac.nz/ml/weka/)*
- *MATLAB® (www.matworks.com)*

يقوم ليو ويسبيرج (*Liu and Weisberg, 2005*) بتطبيق خوارزمية خريطة التنظيم الذاتي (*SOM*) وذلك لغرض تحليل تقلبات المحيط الحالية. كما يقوم يي (*Ye, 2003, Chapter 3*) بتطبيق خريطة التنظيم الذاتي (*SOM*) على بيانات أنشطة الدماغ الخاصة بالقرود وعلاقة ذلك باتجاهات حركتها.

التمارين (Exercises):

١٠-١ واصل عملية التعلُّم في المثال ١٠-١ لعمل تحديثات الوزن، عند إدخال x_2 إلى خريطة التنظيم الذاتي (*SOM*).

١٠-٢ استخدم برمجية *Weka* لرسم خريطة التنظيم الذاتي (*SOM*) للمثال ١٠-١.

١٠-٣ عرف خريطة التنظيم الذاتي (*SOM*) ثنائية الأبعاد، ودالة الجوار في المعادلة ١٠-٢ للمثال ١٠-١ وقم بعمل تكرار واحد لتحديث الوزن عند تقديم $x1$ إلى خريطة التنظيم الذاتي (*SOM*).

١٠-٤ استخدام برمجية *Weka* لرسم خريطة التنظيم الذاتي (*SOM*) ثنائية الأبعاد للمثال ١٠-١.

١٠-٥ استخرج خريطة التنظيم الذاتي (*SOM*) أحادية الأبعاد بنفس دالة الجوار في المثال ١٠-١ لمجموعة البيانات الخاصة بالحلقات الدائرية في مكوك الفضاء في الجدول ١-٢. استخدام درجة حرارة الإطلاق *(Launch – Temprature)*، وضغط التحقق من التسرب *(Leak–Check Pressure)* كمتغيرات خاصية، ودالة التطبيع في المعادلة ٧-٤ للحصول على قيم مطبعة لدرجة حرارة الإطلاق وضغط التحقق من التسرب أيضاً.

١١- التوزيعات الاحتمالية للبيانات الأحادية المتغير
Probability Distributions of Univariate Data

يمكن تطبيق خوارزميات التعنقد الموجودة في الفصول من ٨ إلى ١٠ على بيانات ذات متغير واحد أو أكثر من متغيرات الخاصية. إذا كان هناك متغير خاصية واحد فقط، يكون لدينا بيانات أحادية المتغير. وبالنسبة للبيانات أحادية المتغير، فإن التوزيع الاحتمالي لسجلات البيانات لا يُظهِر فقط عناقيد سجلات البيانات، ولكنه يُظهِر أيضاً العديد من الخصائص الأخرى المتعلقة بتوزيع سجلات البيانات. يمكن تحديد العديد من أنماط البيانات المعينة لبيانات أحادية المتغير من خلال أنواع التوزيعات الاحتمالية المقابلة لها. يستعرض هذا الفصل مفهوم وخصائص التوزيع الاحتمالي، واستخدام خصائص التوزيع الاحتمالي لتحديد بعض أنماط البيانات الأحادية المتغير. وترد قائمة من حزم البرمجيات لتحديد خصائص التوزيع الاحتمالي للبيانات الأحادية المتغير بالإضافة إلى ذكر بعض المراجع لتطبيقات التوزيعات الاحتمالية.

١١-١ التوزيع الاحتمالي للبيانات الأحادية المتغير وخصائص التوزيع الاحتمالي لأنماط بيانات متنوعة

(Probability Distribution of Univariate Data and Probability Distribution Characteristics of Various Data Patterns):

إذا كان لدينا متغير خاصية x وبياناتها المرصودة، $x_1 ,..., x_n$، فإنه غالباً ما يتم استخدام المدرّج التكراري (*frequency histogram*) للبيانات المرصودة بغرض إظهار تكرارات جميع قيم x. يوضح الجدول ١١-١ جميع قيم درجة حرارة الإطلاق (*Launch Temperature*) في مجموعة بيانات الحلقات الدائرية لمكوك الفضاء، والمأخوذة من الجدول ١-٢. ويوضح الشكل ١١-١ مدرجاً تكرارياً لقيم درجة حرارة الإطلاق في الجدول ١١-١ باستخدام عرض فترة يساوي 5 وحدات. إنَّ تغيير عرض الفترة يؤدي إلى تغير تكرارات البيانات المرصودة في كل فترة زمنية، ومن ثم يتبعه تغيير في المدرج التكراري.

في المدرّج التكراري الموضح في الشكل ١١-١، يمكن استبدال المدرج التكراري للبيانات المرصودة لكل فترة زمنية بالكثافة الاحتمالية (*probability density*)، والتي يمكن تقديرها باستخدام نسبة ذلك التكرار إلى العدد الإجمالي لسجلات البيانات المرصودة. من

خلال رسم منحنى ملائم للمدرج التكراري الخاص بالكثافة الاحتمالية، نحصل على منحنى ملائم لدالة الكثافة الاحتمالية *f(x)* التي تعطي الكثافة الاحتمالية لأي قيمة x وهناك نوع شائع من التوزيع الاحتمالي وهو التوزيع الطبيعي (*normal distribution*) بدالة الكثافة الاحتمالية التالية:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{١-١١}$$

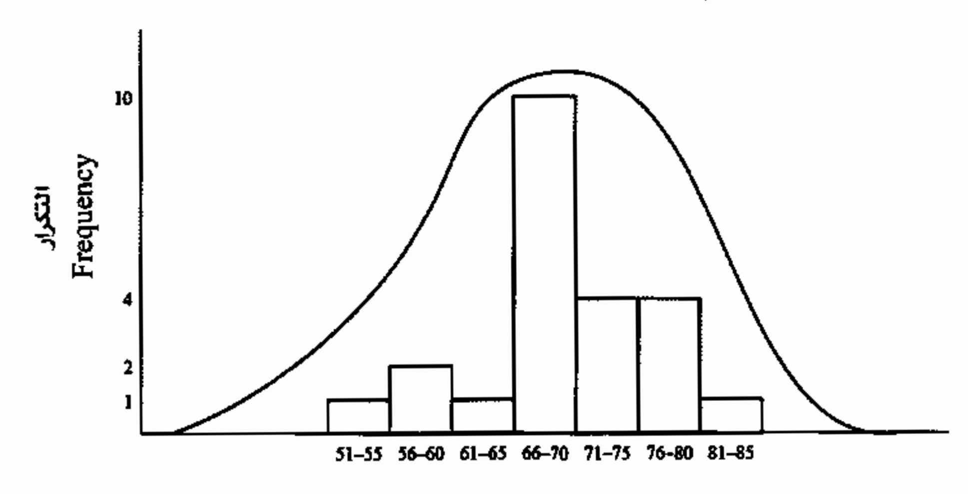
الجدول (١١-١)

قيم درجة حرارة الإطلاق (*Launch Temperature*) في مجموعة البيانات الخاصة بعدد الحلقات الدائرية في مكوك الفضاء

درجة حرارة الاطلاق **Launch Temperature**	رقم الحالة **Instance**
66	1
70	2
69	3
68	4
67	5
72	6
73	7
70	8
57	9
63	10
70	11
78	12
67	13
53	14
67	15
75	16
70	17
81	18
76	19
79	20
75	21
76	22
58	23

الشكل (١١-١)

المدرج التكراري لبيانات درجة حرارة الإطلاق (***Launch Temperature***)



حيث إن:

μ هو المتوسط.

σ هو الانحراف المعياري.

يكون التوزيع الطبيعي متماثلاً مع أعلى كثافة احتمالية عندما يكون المتوسط $x = \mu$ ونفس الكثافة الاحتمالية عند $x= \mu +a$ ، و $x= \mu -a$.

تُظهِر العديد من أنماط البيانات خصائص مميزةً لتوزيعاتها الاحتمالية. على سبيل المثال، درسنا بيانات سلاسل الزمن (*Time series data*) لأنشطة الحاسوب وشبكة الإنترنت (*Ye, 2008, Chapter 9*). تتكون بيانات سلاسل الزمن من بيانات مرصودة على مدى زمني معين. لاحظنا أنماط البيانات التالية المستخرجة من بيانات الحاسوب وشبكة الإنترنت والموضحة في الشكل ١١-٢:

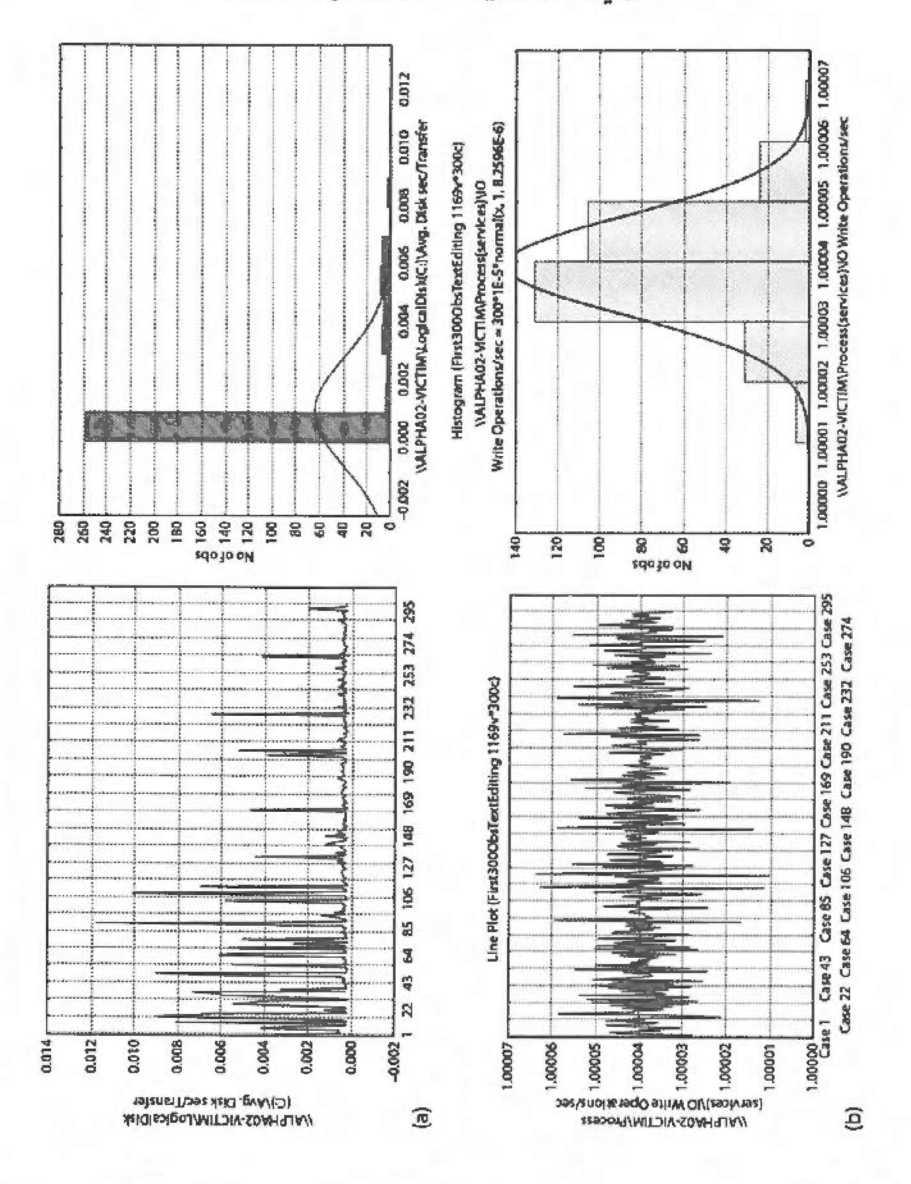
- النمط المسماري (*Spike*).
- نمط التذبذب العشوائي (*Random fluctuation*).
- نمط تغيير الخطوة (*Step change*).
- نمط التغير الثابت (*Steady change*).

هناك خصائص (أو سمَات) خاصة للتوزيعات الاحتمالية لبيانات سلاسل الزمن ذات النمط المسماري، ونمط التّذبذب العشوائي، ونمط تغيير الخطوة، ونمط التغيير الثابت. إن بيانات سلاسل الزمن ذات النمط المسماري كما هو مبين في الشكل ١١-٢ *(a)* ، يكون بها غالبية سجلات البيانات ذات قيم متشابهة، وقليل من سجلات البيانات ذات قيم أعلى، مما ينتج ارتفاعاً مسمارياً تصاعدياً، أو ذات قيم أقل مما ينتج انخفاضاً مسمارياً تنازلياً. يحدد التكرار العالي لسجلات البيانات ذات القيم المتشابهة أين يقع المتوسط ذو الكثافة الاحتمالية العالية، وينتج عن عدد قليل من سجلات البيانات ذات قيم أقل (أو أعلى) من المتوسط لاتجاه مسماري هابط (أو صاعد) ذيل طويل على الجهة اليسرى (أو اليمنى) من المتوسط، ومن ثم توزيع ملتوي (*skewed distribution*) إلى الجهة اليسرى (أو اليمنى). ومن ثم، ينتج عن بيانات سلاسل الزمن المسمارية، توزيع احتمالي ملتوٍ (*skewed probability distribution*)، غير متماثل مع معظم سجلات البيانات التي لها قيم قريبة من المتوسط، وعدد قليل من سجلات البيانات التي لها قيم تنتشر على جانب واحد من المتوسط، والتي تشكل ذيلاً طويلاً، كما هو مبين في الشكل ١١-٢ *(a)*. وينتج عن بيانات سلاسل الزمن ذات نمط التذبذب العشوائي (*random fluctuation*)، توزيع طبيعي، متماثل، كما هو مبين في الشكل ١١-٢ *(b)*. في حين أن بيانات سلاسل الزمن ذات تغيير الخطوة الواحدة (*one step change*)، كما هو مبين في الشكل ١١-٢ *(c)* ، تنتج عنقودين من سجلات البيانات بمركزين متوسطين (*two centroids*) مختلفين، و تنتج من ثم توزيعاً ثنائي النسق (*bimodal distribution*). تقوم بيانات سلاسل الزمن ذات نمط تغييرات الخطوات المتعددة (*multiple step changes*) بإنشاء عناقيد متعددة من سجلات البيانات بمراكز متوسطة مختلفة، ومن ثم إنشاء توزيع متعدد النسق (*multimodal distribution*). ويكون لبيانات السلاسل الزمنية ذات نمط التغيير الثابت (على سبيل المثال: الزيادة الثابتة للقيم أو الانخفاض الثابت للقيم) قيم موزعة بالتساوي، ومن ثم ينتج توزيعاً موحداً، كما هو مبين في الشكل ١١-٢ *(d)*. ولذلك، تنتج الأنماط الأربعة من بيانات سلاسل الزمن أربعة أنواع مختلفة من التوزيع الاحتمالي:

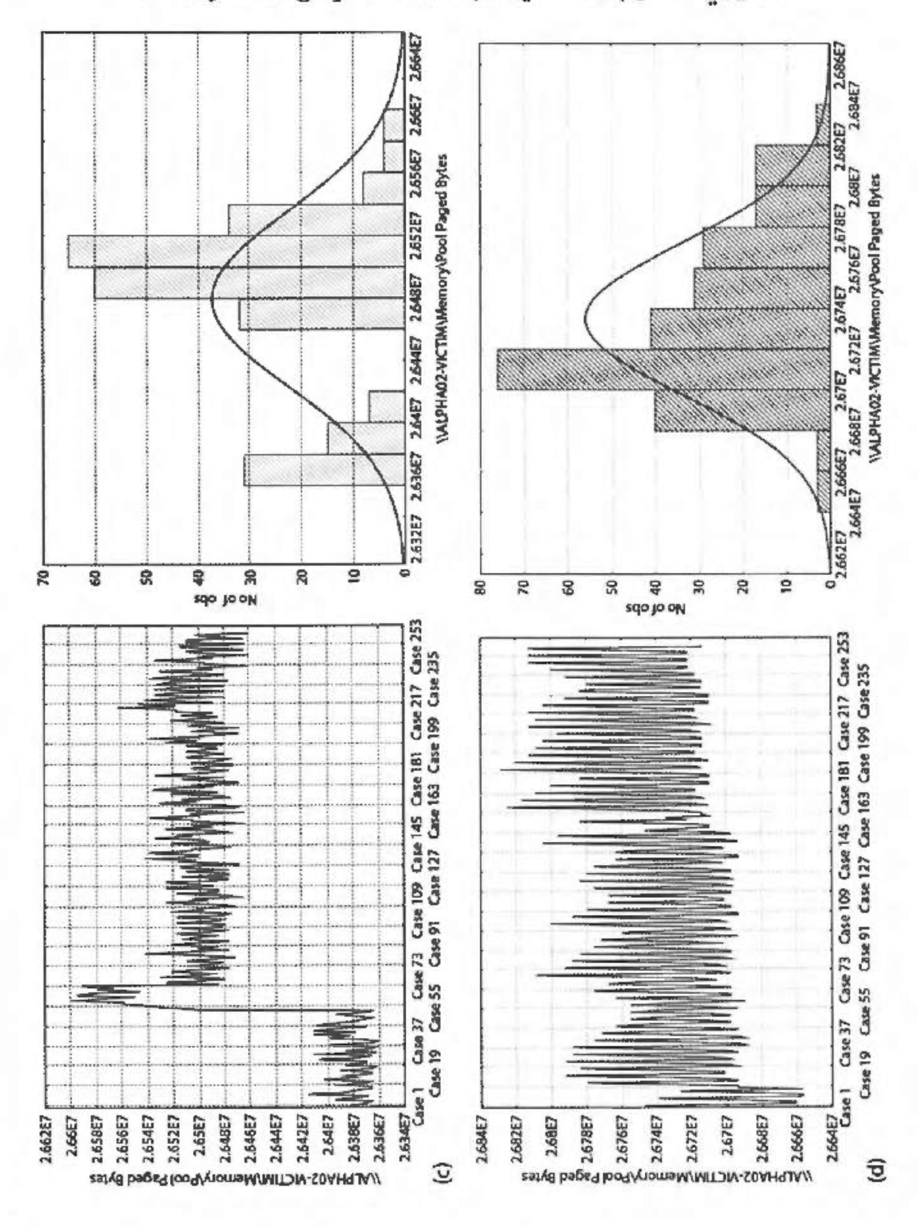
- التوزيع الملتوي الأيمن أو الأيسر *(Left or right skewed distribution)*.
- التوزيع الطبيعي (*Normal distribution*).
- التوزيع المتعدد النسق *(Multimodal distribution)*.
- التوزيع الموحد (*Uniform distribution*).

الشكل (١١-٢)

أنماط بيانات السلاسل الزمنية وتوزيعاتها الاحتمالية. (a) الرسم البياني والمدرّج التكراري الخاص بالنمط المسماري (*spike pattern*)، (b) الرسم البياني والمدرّج التكراري الخاص بنمط التذبذب العشوائي (*randomfluctuation pattern*)



تابع الشكل (١١-٢) أنماط بيانات السلاسل الزمنية وتوزيعاتها الاحتمالية. (c) الرسم البياني والمدرج التكراري الخاص بنمط التغيير بخطوة (*step change pattern*)، (d) الرسم البياني والمدرج التكراري الخاص بنمط التغير الثابت (*steady change pattern*)



كما هو موضح في يي (*Ye, 2008, Chapter 9*)، فإن أنماط البيانات الأربعة، والتوزيعات الاحتمالية المقابلة لها، يمكن استخدامها لتحديد ما إذا كان هناك أنشطة هجومية تجري في أنظمة الحاسوب وعلى شبكة الإنترنت، وذلك لأن بيانات الحاسوب وشبكة الإنترنت التي تتعرض للهجوم، أو لظروف الاستخدام العادي، قد تُظهِر أنماطاً مختلفةً من البيانات. إنَّ الكشف عن الهجمات الإلكترونية يمثل جزءاً مهماً من حماية أنظمة الحاسوب وشبكة الإنترنت من الهجمات الإلكترونية.

١١-٢ طريقة التمييز بين أربعة توزيعات احتمالية

(Method of Distinguishing Four Probability Distribution):

قد نميز أنماط البيانات الأربعة هذه عن طريق تحديد التوزيع الاحتمالي للبيانات الخاصة بها. على الرغم من وجود اختبارات متعددة لتحديد ما إذا كان للبيانات توزيع طبيعيٍ أم لا (*Bryc, 1995*)، فإن الاختبارات الإحصائية لتحديد أحد التوزيعات الاحتمالية لا تُعتبر شائعة. وعلى الرغم من أن المدرج التكراري يمكن رسمه لكي يتيح لنا أولاً أن نتصور، ومن ثمّ نحدد التوزيع الاحتمالي، نحتاج إلى اختبار يمكن برمجته وتشغيله على الحاسوب دون الحاجة إلى الفحص اليدوي والبصري، وخصوصاً عندما تكون مجموعة البيانات كبيرة، وتكون مراقبة البيانات بشكل مباشر مطلوبة مثل التطبيق الخاص بكشف الهجمات الإلكترونية. تمّ تطوير طريقة لتمييز التوزيعات الاحتمالية الأربعة باستخدام خليط عن اختبارات الالتواء أو الانحراف (*skewness*) واختبارات النسق (*mode tests*) في يي (*Ye, 2008, Chapter 9*) والموضح في الجزء التالي.

وتعتمد طريقة تمييز التوزيعات الاحتمالية الأربعة على اختبارات الانحراف والنسق. يتم تعريف الانحراف على أنه:

$$\text{skewness} = E\left(\frac{(x-\mu)^3}{\sigma^3}\right), \qquad (٢\text{-}١١)$$

حيث μ، وσ هما المتوسط والانحراف المعياري لمجتمع البيانات المستهدف للمتغير x عندما يكون لدينا n من سجلات البيانات، x_n...، x_1 فإن انحراف العينة يتم حسابه كما يلي:

$$\text{skewness} = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)(n-2)\, s^3}, \qquad (٣\text{-}١١)$$

حيث $\bar{x}$ وs هما المتوسط والانحراف المعياري لعينة البيانات. وعلى عكس التباين *(variance)*، والذي يقوم بتربيع كل من الانحرافات الموجبة والسالبة عن المتوسط لجعل كل من الانحرافات الموجبة والسالبة عن المتوسط تسهم في التباين بنفس الطريقة، يقوم الانحراف بقياس القدر الذي تكون به انحرافات البيانات عن المتوسط متماثلة ومتطابقة على جانبي المتوسط. يكون للتوزيع المنحرف إلى اليسار بذيل طويل على الجانب الأيسر من المتوسط، قيمةٌ سالبةٌ لمقياس الانحراف. ويكون للتوزيع المنحرف إلى اليمين بذيل طويل على الجانب الأيمن من المتوسط، قيمةٌ موجبةٌ لمقياس الانحراف.

الجدول (١١-٢)

خليط من نتائج اختبارات الانحراف *(Skewness)* والنسق *(Mode)* لتمييز التوزيعات الاحتمالية الأربعة

Skewness Test اختبار الانحراف	Mode Test اختبار النسق	Dip Test اختبار أحادي النسق	Probability Distribution التوزيع الاحتمالي
Any result أي نتيجة	Number of Significant modes ≥ 2 عدد الأنساق ذات الدلالة ≤ ٢	Unimodality is rejected حادية النسق مرفوضةH	Multimodal distribution التوزيع متعدد النسق
Symmetric متماثل	Number of Significant modes > 2 عدد الأنساق ذات الدلالة < ٢	Unimodality is not rejected أحادية النسق غير مرفوضة	Uniform distribution التوزيع الموحد
Symmetric متماثل	Number of Significant modes < 2 عدد الأنساق ذات الدلالة ≤ ٢	Unimodality is not rejected أحادية النسق غير مرفوضة	Normal distribution التوزيع الطبيعي
Skewed منحرف	Number of Significant modes < 2 عدد الأنساق ذات الدلالة ≤ ٢	Unimodality is not rejected أحادية النسق غير مرفوضة	Skewed distribution التوزيع المنحرف

يقع النسق الخاص بالتوزيع الاحتمالي للمتغير x داخل قيمة x التي يكون لها الحد الأقصى من الكثافة الاحتمالية. عندما يكون لدالة الكثافة الاحتمالية قيم قصوى متعددة محلية (*multiple local maxima*)، يكون للتوزيع الاحتمالي أنساق (*modes*) متعددة. الكثافة الاحتمالية ذات القيمة الكبيرة تشير إلى عنقود من سجلات البيانات المتشابهة. ومن ثم، يرتبط النسق بعملية تعنقد سجلات البيانات. التوزيع الطبيعي (*normal distribution*)، والتوزيع المنحرف (*skewed distribution*)، هي أمثلة على التوزيعات أحادية النسق الواحد (*unimodal distributions*) ، وذلك على العكس من التوزيعات المتعددة النسق (*distributions multimodal*) ذات الأنساق المتعددة. التوزيع الموحد *(uniform distribution)* ليس له نسقٌ ذو دلالة مهمة، وذلك لأن البيانات موزعة بشكلٍ متساوٍ، ولا تتشكل في عناقيد. يحدد اختبار أحادي النسق (*dip test*) (*Hartigan and Hartigan, 1985*) ما إذا كان، التوزيع الاحتمالي الأحادي النسق. يحدد اختبار النسق في البرنامج الإحصائي *R* (*www.r-project.org*) الدلالة المهمة لكل نسق محتمل في التوزيع الاحتمالي، ويعطي عدد الأنساق ذات الدلالة المهمة.

يوضح الجدول ١١-٢ خليطاً من نتائج اختبارات الانحراف والنسق والتي تُستخدَم لتمييز التوزيعات الاحتمالية الأربعة: التوزيع متعدد النسق (*multimodal distribution*) بما فيها التوزيع ثنائي النسق (*bimodal distribution*)، والتوزيع الموحد (*uniform distribution*)، والتوزيع الطبيعي (*normal distribution*)، والتوزيع المنحرف (*skewed distribution*). لذلك، إذا علمنا أن للبيانات واحداً من هذه التوزيعات الاحتمالية الأربعة، يمكننا التحقق من خليط النتائج المكون من اختبار أحادي النسق (*dip test*)، واختبار النسق (*mode test*)، واختبار الانحراف (*skewness test*)، وتحديد أي من التوزيعات الاحتمالية تحمله البيانات.

١١-٣ البرمجيات والتطبيقات (Software and Applications):

يقوم برنامج ستاتسيكا *(Statistica) (www.statsoft.com)*، بدعم اختبار الانحراف (*skewness test*). وتدعم برامج *R* الإحصائي (*www.r-project.org*، *www.cran.r-project.org/doc/packages/diptest.pdf*) اختبار أحادي النسق (*dip test*)، واختبار النسق (*mode test*). في يي (*Ye, 2008, Chapter 9*)، يمكن تمييز

البيانات الحاسوبية، وبيانات شبكة الإنترنت التي تتعرض للهجوم الإلكتروني، وظروف الاستخدام الطبيعي، وذلك عن طريق التوزيعات الاحتمالية المختلفة للبيانات في ظل ظروف مختلفة.

يتم إجراء الكشف عند التعرض للهجوم عبر الإنترنت من خلال مراقبة البيانات الحاسوبية المرصودة، وبيانات شبكة الإنترنت، وتحديد ما إذا كان التغيير على التوزيع الاحتمالي من وضع الاستخدام الطبيعي إلى وضع الهجوم الإلكتروني قد حدث أم لا.

التمارين (Exercises):

١١-١ قم باختيار واستخدام البرمجية لإجراء اختبار الانحراف، واختبار النسق، والاختبار أحادي النسق، لبيانات درجة حرارة الإطلاق *(Launch Temperature)* في الجدول ١١-١، وقم باستخدام نتائج الاختبار لتحديد ما إذا كان التوزيع الاحتمالي لبيانات درجة حرارة الإطلاق يقع في أحد التوزيعات الاحتمالية الأربعة في الجدول ١١-٢.

١١-٢ اختر متغيراً رقمياً في مجموعة البيانات التي حصلت عليها في المسألة رقم ١-٢ وقم باختيار عرض الفترة لرسم مدرج تكراري للبيانات الخاصة بالمتغير. قم باختيار واستخدام البرمجية لإجراء اختبار الانحراف، واختبار النسق، واختبار أحادي النسق، على البيانات الخاصة بالمتغير، واستخدم نتائج الاختبار لتحديد ما إذا كان التوزيع الاحتمالي لبيانات درجة حرارة الإطلاق يقع في واحد من التوزيعات الاحتمالية الأربعة في الجدول ١١-٢.

١١-٣ اختر متغيراً رقمياً في مجموعة البيانات التي حصلت عليها في المسألة ١-٣، وقم باختيار عرض الفترة لرسم مدرج تكراري للبيانات الخاصة بالمتغير. قم باختيار واستخدام البرنامج لإجراء اختبار الانحراف، واختبار النسق، واختبار أحادي النسق، على البيانات الخاصة بالمتغير، وقم باستخدام نتائج الاختبار لتحديد ما إذا كان التوزيع الاحتمالي لبيانات درجة حرارة الإطلاق يقع في واحد من التوزيعات الاحتمالية الأربعة في الجدول ١١-٢.

١٢- قواعد الاقتران
Association Rules

تكشف قواعدُ الاقتران (*association rules*) العناصرَ (*items*) التي كثيراً ما يرتبط بعضها ببعض. لقد تمّ تطوير خوارزمية قواعد الاقتران بدايةً في سياق تحليل سلة السوق (*market basket analysis*) لدراسة السلوكيات الشرائية للعملاء والتي يمكن استخدامها لغرض التسويق. تكشف قواعد الاقتران ما هي العناصر التي غالباً ما يشتريها العملاء معاً. إن العناصر أو المواد التي، في كثير من الأحيان، يتم شراؤها معاً يمكن وضعها في المتاجر أو يمكن أن يتم ربطها معاً في مواقع التجارة الإلكترونية على الإنترنت لتعزيز مبيعات هذه المواد أو لأغراض تسويقية أخرى. يوجد العديد من التطبيقات الأخرى لقواعد الاقتران، على سبيل المثال، تحليل النصوص (*text analysis*) لغرض تصنيف الوثائق واسترجاعها. يقدم هذا الفصل خوارزمية استكشاف قواعد الاقتران. وترد قائمة بحزم البرمجيات التي تدعم قواعد الاقتران. ويتم إعطاء بعض التطبيقات لقواعد الاقتران مع مراجعها.

١-١٢ تعريف قواعد الاقتران ومقاييس الاقتران
(Definition of Association Rules and Measures of Association):

تحتوي مجموعة العناصر (*item set*) على مجموعة من العناصر. على سبيل المثال، تعد عملية شراء عميل في متجر ما (بقالة) هي مجموعة عناصر أو مجموعة من مواد البقالة مثل البيض والطماطم والتفاح. تحتوي مجموعة البيانات لاكتشاف أعطال النظام بتسع حالات من الأعطال الآلية الأحادية في الجدول ٨-١ على تسعة سجلات للبيانات، والتي يمكن اعتبارها تسع مجموعات من العناصر عن طريق أخذ x_1، x_2، x_3، x_4، x_5، x_6، x_7، x_8، x_9 كتسع مشكلات جودة مختلفة وبقيمة تساوي 1 والتي تشير إلى وجود مشكلة جودة. ويوضح الجدول ١٢-١ مجموعات العناصر التسع التي تمّ الحصول عليها من مجموعة بيانات اكتشاف أعطال النظام. ويكشف اقتران العناصر المتكرر في الجدول ١٢-١ عن أي من مشاكل الجودة والتي غالباً ما تحدث معاً.

وتأخذ قاعدة الاقتران الشكل:

$$A \rightarrow C$$

حيث إن:

A هي مجموعة عناصر وتُسمى الشرط السابق (*antecedent*).
C هي مجموعة عناصر وتُسمى النتيجة اللاحقة (*consequent*).

A وC ليس لديهما أي عناصر مشتركة، وهذا يعني أن، $A \cap C = \emptyset$ (مجموعة فاي). إن العلاقة بين A وC في قاعدة الاقتران تعني أن وجود مجموعة عناصر A في سجل بيانات تعني وجود مجموعة العناصر C في سجل البيانات نفسه، وهذا يعني أن مجموعة العناصر C مقترنة بمجموعة العنصر A.

الجدول (١٢-١)

مجموعة بيانات اكتشاف أعطال النظام بتسع حالات من الأعطال الآلية الأحادية ومجموعات العنصر التي تمّ الحصول عليها من مجموعة البيانات هذه

رقم الحالة - Instance (الآلة المعطلة - Faulty Machine)	متغيرات الخاصية عن جودة وحدات المنتج Attribute Variables about Quality of Parts									العناصر في كل سجل بيانات Items in Each Data Record
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	
1 (M1)	1	0	0	0	1	0	1	0	1	$\{x_1, x_5, x_7, x_9\}$
2(M2)	0	1	0	1	0	0	0	1	0	$\{x_2, x_4, x_8\}$
3(M3)	0	0	1	1	0	1	1	1	0	$\{x_3, x_4, x_6, x_7, x_8\}$
4(M4)	0	0	0	1	0	0	0	1	0	$\{x_4, x_8\}$
5(M5)	0	0	0	0	1	0	1	0	1	$\{x_5, x_7, x_9\}$
6(M6)	0	0	0	0	0	1	1	0	0	$\{x_6, x_7\}$
7(M7)	0	0	0	0	0	0	1	0	0	$\{x_7\}$
8(M8)	0	0	0	0	0	0	0	1	0	$\{x_8\}$
9(M9)	0	0	0	0	0	0	0	0	1	$\{x_9\}$

يتم تعريف مقاييس الدعم (*support*)، الثقة (*confidence*)، والعون (*lift*) واستخدمها لاكتشاف مجموعتي العناصر A وC اللتين كثيراً ما تقترنان معاً. مقياس الدعم أو *support (x)* في مجموعة العناصر X يقيس نسبة سجلات البيانات التي تحتوي على مجموعة العناصر x ويعرف بأنه:

$$support(X) = \frac{|\{S|S \in D \ and \ S \supseteq X\}|}{N}, \qquad (١-١٢)$$

حيث إن:

D يدل على مجموعة البيانات التي تحتوي على سجلات البيانات.

S هو سجل بيانات في D (المشار إليه بـ $S \in D$) ويحتوي على العناصر في X (المشار إليها بـ $S \supseteq X$).

$|\ |$ تدل على عدد سجلات البيانات في S.

N هو عدد سجلات البيانات في D.

استناداً إلى التعريف، يكون لدينا:

$$support(\emptyset) = \frac{|\{S|S \in D \ and \ S \supseteq \emptyset\}|}{N} = \frac{N}{N} = 1.$$

على سبيل المثال، لمجموعة البيانات التي لها تسعة سجلات بيانات في الجدول ١٢-١،

$$support(\{x_5\}) = \frac{2}{9} = 0.22$$
$$support(\{x_7\}) = \frac{5}{9} = 0.56$$
$$support(\{x_9\}) = \frac{3}{9} = 0.33$$
$$support(\{x_5, x_7\}) = \frac{2}{9} = 0.22$$
$$support(\{x_5, x_9\}) = \frac{2}{9} = 0.22.$$

مقياس الدعم ، أو $support\ (A{\rightarrow}C)$، يقيس نسبة سجلات البيانات التي تحتوي على كل من الشرط السابق A والنتيجة اللاحقة C في قاعدة الاقتران $A{\rightarrow}C$، ويعرف بأنه:

$$support(A \rightarrow C) = support(A \cup C)\ , \qquad (٢-١٢)$$

حيث $A \cup C$ عبارة عن اتحاد لمجموعة العناصر A ومجموعة العناصر C وتحتوي على عناصر من A وC. استناداً إلى التعريف، يكون لدينا:

$$support(\emptyset \rightarrow C) = support(C)$$
$$support(A \rightarrow \emptyset) = support(A).$$

على سبيل المثال:

$$support(\{x_5\} \rightarrow \{x_7\}) = support(\{x_5\} \cup \{x_7\}) = support(\{x_5, x_7\}) = 0.22$$

$$support(\{x_5\} \rightarrow \{x_9\}) = support(\{x_5\} \cup \{x_9\}) = support(\{x_5, x_9\}) = 0.22.$$

مقياس الثقة، أو $confidence\ (A{\rightarrow}C)$، يقيس نسبة سجلات البيانات المحتوية على الشرط السابق A والتي بدورها أيضاً تحتوي على النتيجة اللاحقة C، ويعرف بأنه:

$$confidence(A \rightarrow C) = \frac{support(A \cup C)}{support(A)}\ . \qquad (٣-١٢)$$

استناداً إلى التعريف، يكون لدينا:

$$confidence(\emptyset \to C) = \frac{support(C)}{support(\emptyset)} = \frac{support(C)}{1} = support(C)$$

$$confidence(A \to \emptyset) = \frac{support(A)}{support(A)} = 1.$$

على سبيل المثال:

$$confidence(\{x_5\} \to \{x_7\}) = \frac{support(\{x_5\} \cup \{x_7\})}{support(\{x_5\})} = \frac{0.22}{0.22} = 1$$

$$confidence(\{x_5\} \to \{x_9\}) = \frac{support(\{x_5\} \cup \{x_9\})}{support(\{x_5\})} = \frac{0.22}{0.22} = 1.$$

إذا كان الشرط السابق A والنتيجة اللاحقة C مستقلتين عن بعضهما و*support (C)* له قيمة عالية (وهو ما يعني وجود النتيجة اللاحقة في العديد من سجلات البيانات في مجموعة البيانات)، فإن *support* $(A \cup C)$ سيكون له قيمة عالية لأن C موجودة في العديد من سجلات البيانات التي تحتوي أيضاً على A. ونتيجة لذلك، نحصل على قيمة عالية لـ *support* $(A \to C)$ و *confidence* $(A \to C)$ على الرغم من كون A وC مستقلتين عن بعضهما واقتران $A \to C$ يكون له فائدة قليلة. على سبيل المثال، إذا تمّ احتواء مجموعة العناصر C في كل سجل بيانات في مجموعة البيانات، يكون لدينا:

$$support(A \to C) = support(A \cup C) = support(A)$$

$$confidence(A \to C) = \frac{support(A \cup C)}{support(A)} = \frac{support(A)}{support(A)} = 1.$$

لكن، تظل قاعدة اقتران $A \rightarrow C$ ذات فائدة قليلة بالنسبة لنا، لأن مجموعة العناصر C موجودة في كل سجل بيانات، ومن ثم فإن أي مجموعة عناصر بما في ذلك A تقترن مع C. ولمعالجة هذه المسألة، يتم تعريف مقياس العون، أو *lift* $(A \rightarrow C)$، على أنه:

$$lift(A \rightarrow C) = \frac{confidence(A \rightarrow C)}{support(C)} = \frac{support(A \cup C)}{support(A) \times support(C)}. \qquad (٤\text{-}١٢)$$

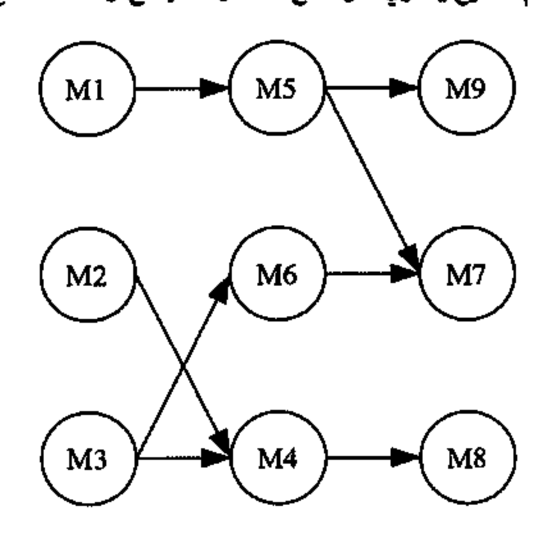
إذا كان الشرط السابق A والنتيجة اللاحقة C مستقليتين عن بعضهما ولكن الدعم (C) ***support*** له قيمة مرتفعة، فإن هذه القيمة المرتفعة تعطي قيمة منخفضة لـ ***lift*** $(A \rightarrow C)$. على سبيل المثال:

$$lift(\{x_5\} \rightarrow \{x_7\}) = \frac{confidence(\{x_5\} \rightarrow \{x_7\})}{support(\{x_7\})} = \frac{1}{0.56} = 1.79$$

$$lift(\{x_5\} \rightarrow \{x_9\}) = \frac{confidence(\{x_5\} \rightarrow \{x_9\})}{support(\{x_9\})} = \frac{1}{0.33} = 3.03.$$

الشكل (١٢-١)

نظام تصنيع يحتوي على تسع آلات وخط إنتاج وحدات المنتج



يكون لقواعد الاقتران، $\{x_7\} \longrightarrow \{x_5\}$ و$\{x_9\} \longrightarrow \{x_5\}$ نفس قيم مقياس الدعم *(support)* والثقة *(confidence)* ولكن قيم مختلفة لمقياس العون *(lift)*. ومن ثم، يظهر أن x_5 يكون لها تأثير أكبر على تكرار x_9 أكثر من تكرار x_7. الشكل ١-١، الذي يتم نسخه في الشكل ١٢-١، يعطي تدفقات وخط إنتاج وحدات المنتج لمجموعة البيانات في الجدول ١٢-١. كما هو مبين في الشكل ١٢-١، تذهب وحدات المنتج التي تتدفق من خلال الآلة الخامسة *M5* إلى الآلة السابعة *M7* والآلة التاسعة *M9*. ومن ثم، ينبغي أن يكون لـ x_5 نفس التأثير على x_7 وx_9. لكن، وحدات المنتج المتدفقة خلال الآلة السادسة *M6* تذهب أيضاً إلى الآلة السابعة *M7*، وتكون x_7 أكثر تكراراً من x_9 في مجموعة البيانات، مما ينتج عنه قيمة عون *(lift)* أقل لـ $\{x_7\} \longrightarrow \{x_5\}$ من تلك لـ $\{x_9\} \longrightarrow \{x_5\}$. وبعبارة أخرى، فإن x_7 لا تتأثر بـ x_5 فحسب، بل أيضاً بـ x_6 و x_3 كما هو مبين في الشكل ١٢-١، مما يجعل x_7 تظهر أقل اعتماداً على x_5 لأن مقياس العون *(lift)* يعالج مسألة استقلالية كل من الشرط السابق والنتيجة اللاحقة من خلال قيمة عون منخفضة.

١٢-٢ اكتشاف قاعدة الاقتران (Association Rule Discovery):

يُستخدم اكتشاف قاعدة الاقتران (*association rule discovery*) للعثور على جميع قواعد الاقتران التي تتجاوز الحد الأدنى للحدود (*thresholds*) في مقاييس معينة للاقتران، وعادةً ما تكون مقاييس الدعم *(support)* والثقة (*confidence*). يتم بناء قواعد الاقتران باستخدام مجموعات عناصر متكررة التي تحقق الحد الأدنى من الدعم. بإعطاء مجموعة بيانات من سجلات البيانات المكونة من عدد p من العناصر كحد أقصى، فمن شأن مجموعة العناصر أن تكون ممثلةً على النحو التالي $(x_1, \ldots, x_p)$، $x_i = 0$ أو $x_i = 1$ ،$i = 1, \ldots, p$، حيث إن $x_i = 1$ تشير إلى وجود العنصر رقم i في مجموعة العناصر. وبما أن هناك عدد 2^p من التركيبات الممكنة للقيم المختلفة لـ $(x_i, \ldots, x_p)$، فهناك مجموعات عناصر مختلفة وممكنة عددها $(2^p - 1)$ لعدد 1 إلى p من العناصر، ماعدا المجموعة الفارغة الممثلة بـ $(0, \ldots, 0)$. ومن غير العملي القيام بفحص شامل لقيمة مقياس الدعم (*support*) لكل واحد من مجموعات العناصر المختلفة الممكنة $(2^p - 1)$.

تقدم خوارزمية أبريوري (الأسبقية) (*Apriori algorithm*) (*Agrawal and Srikant, 1994*) إجراء فعّالاً لتوليد مجموعات العناصر المتكررة من خلال الأخذ في

الاعتبار أن مجموعة العناصر لا يمكن أن تكون مجموعة عناصر متكررة إلا إذا كانت جميع المجموعات الفرعية منها هي مجموعات عناصر متكررة. يوضح الجدول ١٢-٢ خطوات خوارزمية أبريوري (الأسبقية) لمجموعة بيانات محددة D.

في الخطوة ٥ من خوارزمية أبريوري (الأسبقية)، يكون لمجموعتي العناصر من F_{i-1} العناصر نفسها من $x_1,...x_{i-2}$ وتختلف مجموعتا العناصر فقط في عنصر واحد بكون x_{i-1} موجودةً في مجموعة عنصر واحد و x_i موجودة في مجموعة عناصر آخرى. يتم بناء مجموعة عناصر مرشحة لـ F_i من خلال إدراج $x_1 ,...x_{i-2}$ (العناصر المشتركة لمجموعتي العنصر من F_{i-1})، و x_{i-1} و x_i على سبيل المثال، إذا كانت $\{x_1 ،x_2 ، x_3\}$ هي مجموعة متكررة بثلاث عناصر ، فإن أي تشكيل مكون من عنصرين من هذه المجموعة المتكررة ، $\{x_1 ،x_2\}$، $\{x_1 ، x_3\}$، أو $\{x_2 ، x_3\}$، يجب أن يكون مجموعة متكررة بعنصرين. وهذا يعني أنه إذا كان $support (\{x_1 ،x_2 ، x_3\})$ أكبر من أو يساوي الحد الأدنى للدعم، فـإن $support (\{x_1 ، x_2\})$، $support (\{x_2 ، x_3\})$ ، $support (\{x_1 ، x_3\})$ يجب أن يكون أكبر من أو يساوي الحد الأدنى للدعم. ومن ثم المجموعة المتكررة ذات الثلاث عناصر، $\{x_1 ،x_2 ، x_3\}$، يمكن بناؤها باستخـدام اثنين من مجمـوعاتها الفـرعية ذات العنصـرين والتي تختلف في عنصر واحد فقط، $\{x_1 ، x_2\}$ و$\{x_1 ، x_3\}$، $\{x_1 ، x_2\}$ و$\{x_2 ، x_3\}$، $\{x_1 ، x_3\}$ و$\{x_2 ، x_3\}$. وبالمثل، فإن أي مجموعة متكررة ذات i عنصر يجب أن تأتي من مجموعات متكررة ذات $(i - 1)$ عنصر والتي تختلف في عنصر واحد فقط. تقلل هذه الطريقة لبناء مجموعة عناصر مرشحة لـ F_i، وبدلالة هامة، من عدد مجموعات العناصر المرشحة لـ F_i التي سيتم تقييمها في الخطوة ٧ من الخوارزمية.

يوضح المثال ١٢-١ استخدام خوارزمية أبريوري (الأسبقية). عندما تكون البيانات متناثرة *(sparse)* بحيث يكون كل عنصر غير متكرر نسبياً في مجموعة البيانات، تكون خوارزمية أبريوري (الأسبقية) فعّالة حيث أنها تعطي عدداً صغيراً من مجموعات العناصر المتكررة، بحيث يحتوي عدد قليل منها على أعداد كبيرة من العناصر. وعندما تكون البيانات كثيفة *(dense)*، تكون خوارزمية أبريوري (الأسبقية) أقل كفاءةً وتعطي عدداً كبيراً من مجموعات العناصر المتكررة.

الجدول (٢-١٢)

خوارزمية أبريوري (الأسبقية) (*Apriori Algorithm*) - (إنجليزي وعربي)

Step	Description of the Step
1	F_1 = {frequent one-item sets}
2	$i = 1$
3	while $F_i \neq \varnothing$
4	$i = i + 1$
5	$C_i = \{\{x_1, \ldots, x_{i-2}, x_{i-1}, x_i\} \mid \{x_1, \ldots, x_{i-2}, x_{i-1}\} \in F_{i-1}$ and $\{x_1, \ldots, x_{i-2}, x_i\} \in F_{i-1}\}$
6	for all data records $S \in D$
7	for all candidate sets $C \in C_i$
8	if $S \supseteq C$
9	C.count = C.count + 1
10	F_i = {C | C ∈ C_i and C.count ≥ minimum support}
11	return all F_j, $j = 1, \ldots, i-1$

الخطوة	الوصف
١	لتكن F_1 = { مجموعات متكررة ذات عنصر-واحد}
٢	$i = 1$
٣	كرّر (WHILE) مادام أن $F_i \neq \varnothing$
٤	$i = i + 1$
٥	$C_i=\{\{x_1, \ldots, x_{i-2}, x_{i-1}, x_i\} \mid \{x_1, \ldots, x_{i-2}, x_{i-1}\} \in F_{i-1}$ and $\{x_1, \ldots, x_{i-2}, x_i\} \in F_{i-1}\}$
٦	لكل سجلات البيانات، $S \in D$
٧	لكل مجموعات العناصر المرشحة، $C \in C_i$
٨	إذا (if) كان، $S \supseteq C$
٩	C.count = C.count+1
١٠	F_i = {C | C C_i and C.count ≥ مقياس الدعم الأدنى}
١١	رجّع كل مجموعات العناصر F_j ، حيث $j = 1, \ldots, i-1$

المثال (١٢-١):

من مجموعة البيانات في الجدول ١٢-١، قم بإيجاد كل مجموعات العناصر المتكررة ذات مقياس الدعم بقيمة حد أدنى تساوي*0.2* أو (*min- support= 0.2*). بفحص مقياس الدعم لكل مجموعة عناصر بعنصر واحد، نحصل على:

$$F_1 = \Big\{\{x_4\}, support = \frac{3}{9} = 0.33, \\ \{x_5\}, support = \frac{2}{9} = 0.22, \\ \{x_6\}, support = \frac{2}{9} = 0.22, \\ \{x_7\}, support = \frac{5}{9} = 0.56, \\ \{x_8\}, support = \frac{4}{9} = 0.44 \\ \{x_9\}, support = \frac{3}{9} = 0.33\Big\}.$$

باستخدام مجموعات العناصر المتكررة ذات للعنصر الواحد لتكوين المجموعات المتكررة ذات العنصرين وفحص مقياس دعمهم، نحصل على:

$$F_2 = \Big\{\{x_4, x_8\}, support = \frac{3}{9} = 0.33, \\ \{x_5, x_7\}, support = \frac{2}{9} = 0.22, \\ \{x_5, x_9\}, support = \frac{2}{9} = 0.22, \\ \{x_6, x_7\}, support = \frac{2}{9} = 0.22, \\ \{x_7, x_9\}, support = \frac{2}{9} = 0.22\Big\}.$$

حيث إن {x_7 ، x_9} و{x_5 ، x_9} ،{x_5 ، x_7} تختلف عن بعضها البعض في عنصر واحد فقط، فيتم استخدامهم لبناء مجموعة ذات ثلاثة عناصر {x_5 ،x_7 ، x_9} - مجموعة الثلاث عناصر الوحيدة التي يمكن بناؤها:

$$F_3 = \left\{\{x_5, x_7, x_9\}, \qquad support = \frac{2}{9} = 0.22\right\}.$$

لاحظ أن بناء مجموعة ذات ثلاثة عناصر من مجموعات ذات عنصرين والتي تختلف في أكثر من عنصر واحد لا يعطي مجموعة متكررة ذات ثلاثة عناصر. على سبيل المثال، {x_4 ،x_8} و{x_5 ، x_7} هي مجموعات متكررة ذات عنصرين والتي تختلف في عنصرين. المجموعات {x_4 ، x_5} و{x_4 ، x_7}، {x_8 ، x_5} و{x_8 ، x_7} ليست مجموعات متكررة ذات عنصرين. يتم بناء أي مجموعة بثلاث عناصر باستخدام {x_4 ، x_8} و{x_5 ، x_7}، على سبيل المثال، {x_4 ،x_5 ، x_8}، ليست مجموعة متكررة مكونة من ثلاثة عناصر نظراً لأنه ليس كل زوج بعنصرين مكون من {x_4 ،x_5 ، x_8} هو مجموعة متكررة ذات عنصرين. على وجه التحديد، فإن {x_4 ، x_5} و{x_8 ، x_5} ليست مجموعات متكررة ذات عنصرين.

نظراً لوجود مجموعة واحدة متكررة فقط مكونة من ثلاثة عناصر، فلا يمكننا توليد مجموعة مرشحة مكونة من أربعة عناصر في الخطوة ٥ من خوارزمية أبريوري (الأسبقية). وهذا يعني أن، $C_4 = \emptyset$، ونتيجة لذلك، فان $F_4 = \emptyset$ في الخطوة ٣ من خوارزمية أبريوري (الأسبقية)، ونقوم بالخروج من تعليمة التكرار (*WHILE*). في الخطوة ١١ من خوارزمية أبريوري (الأسبقية)، نقوم بجمع جميع مجموعات العنصر المتكررة التي تحقق *min-support= 0.2*:

{x_4} ،{x_5} ،{x_6} ،{x_7} ،{x_8} ،{x_9} ،{x_4 ، x_8} ،{x_5 ، x_7} ،{x_5 ، x_9} ،{x_6 ، x_7} ،{x_7 ، x_9} ،{x_5 ، x_7 x_9}

المثال (١٢-٢):

قم باستخدام مجموعات العناصر المتكررة من المثال ١٢-١ لتوليد جميع قواعد الاقتران التي تحقق الحد الأدنى للدعم *min- support= 0.2* والحد الأدنى للثقة *min-confidence= 0.5*.

باستخدام كل مجموعة عناصر متكررة F التي تمّ الحصول عليها من المثال ١٢-١، نقوم بتوليد كل من قواعد الاقتران التالية، $A \rightarrow C$، التي تحقق:

$$A \cup C = F,$$
$$A \cap C = \emptyset,$$

معايير الحد الأدنى للدعم *min-support* والحد الأدنى للثقة *min-confidence* :

$$\emptyset \rightarrow \{x_4\}, support = 0.33, confidence = 0.33$$
$$\emptyset \rightarrow \{x_5\}, support = 0.22, confidence = 0.22$$
$$\emptyset \rightarrow \{x_6\}, support = 0.22, confidence = 0.22$$
$$\emptyset \rightarrow \{x_7\}, support = 0.56, confidence = 0.56$$
$$\emptyset \rightarrow \{x_8\}, support = 0.44, confidence = 0.44$$
$$\emptyset \rightarrow \{x_9\}, support = 0.33, confidence = 0.33$$
$$\emptyset \rightarrow \{x_4, x_8\}, support = 0.33, confidence = 0.33$$
$$\emptyset \rightarrow \{x_5, x_7\}, support = 0.22, confidence = 0.22$$
$$\emptyset \rightarrow \{x_5, x_9\}, support = 0.22, confidence = 0.22$$
$$\emptyset \rightarrow \{x_6, x_7\}, support = 0.22, confidence = 0.22$$
$$\emptyset \rightarrow \{x_7, x_9\}, support = 0.22, confidence = 0.22$$
$$\emptyset \rightarrow \{x_5, x_7, x_9\}, support = 0.22, confidence = 0.22$$
$$\{x_4\} \rightarrow \emptyset, support = 0.33, confidence = 1$$
$$\{x_5\} \rightarrow \emptyset, support = 0.22, confidence = 1$$
$$\{x_6\} \rightarrow \emptyset, support = 0.22, confidence = 1$$
$$\{x_7\} \rightarrow \emptyset, support = 0.56, confidence = 1$$
$$\{x_8\} \rightarrow \emptyset, support = 0.44, confidence = 1$$
$$\{x_9\} \rightarrow \emptyset, support = 0.33, confidence = 1$$

$$\{x_4, x_8\} \rightarrow \emptyset, support = 0.33, confidence = 1$$

$$\{x_5, x_7\} \rightarrow \emptyset, support = 0.22, confidence = 1$$

$$\{x_5, x_9\} \rightarrow \emptyset, support = 0.22, confidence = 1$$

$$\{x_6, x_7\} \rightarrow \emptyset, support = 0.22, confidence = 1$$

$$\{x_7, x_9\} \rightarrow \emptyset, support = 0.22, confidence = 1$$

$$\{x_5, x_7, x_9\} \rightarrow \emptyset, support = 0.22, confidence = 1$$

$$\{x_4\} \rightarrow \{x_8\}, support = 0.33, confidence = 1$$

$$\{x_5\} \rightarrow \{x_7\}, support = 0.22, confidence = 1$$

$$\{x_5\} \rightarrow \{x_9\}, support = 0.22, confidence = 1$$

$$\{x_6\} \rightarrow \{x_7\}, support = 0.22, confidence = 1$$

$$\{x_7\} \rightarrow \{x_9\}, support = 0.22, confidence = 0.39$$

$$\{x_8\} \rightarrow \{x_4\}, support = 0.33, confidence = 0.75$$

$$\{x_7\} \rightarrow \{x_5\}, support = 0.22, confidence = 0.39$$

$$\{x_9\} \rightarrow \{x_5\}, support = 0.22, confidence = 0.67$$

$$\{x_7\} \rightarrow \{x_6\}, support = 0.22, confidence = 0.39$$

$$\{x_9\} \rightarrow \{x_7\}, support = 0.22, confidence = 0.67$$

$$\{x_5\} \rightarrow \{x_7, x_9\}, support = 0.22, confidence = 1$$

$$\{x_7\} \rightarrow \{x_5, x_9\}, support = 0.22, confidence = 0.39$$

$$\{x_9\} \rightarrow \{x_5, x_7\}, support = 0.22, confidence = 0.67$$

$$\{x_7, x_9\} \rightarrow \{x_5\}, support = 0.22, confidence = 1$$

$$\{x_5, x_9\} \rightarrow \{x_7\}, support = 0.22, confidence = 1$$

$$\{x_5, x_7\} \rightarrow \{x_9\}, support = 0.22, confidence = 1.$$

بإزالة كل قاعدة اقتران في شكل $F \rightarrow \emptyset$، نحصل على المجموعة النهائية من قواعد الاقتران:

$$\{x_4\} \rightarrow \emptyset, support = 0.33, confidence = 1$$

$$\{x_5\} \rightarrow \emptyset, support = 0.22, confidence = 1$$

$$\{x_6\} \rightarrow \emptyset, support = 0.22, confidence = 1$$

$$\{x_7\} \rightarrow \emptyset, support = 0.56, confidence = 1$$

$$\{x_8\} \to \emptyset, support = 0.44, confidence = 1$$

$$\{x_9\} \to \emptyset, support = 0.33, confidence = 1$$

$$\{x_4, x_8\} \to \emptyset, support = 0.33, confidence = 1$$

$$\{x_5, x_7\} \to \emptyset, support = 0.22, confidence = 1$$

$$\{x_5, x_9\} \to \emptyset, support = 0.22, confidence = 1$$

$$\{x_6, x_7\} \to \emptyset, support = 0.22, confidence = 1$$

$$\{x_7, x_9\} \to \emptyset, support = 0.22, confidence = 1$$

$$\{x_5, x_7, x_9\} \to \emptyset, support = 0.22, confidence = 1$$

$$\{x_4\} \to \{x_8\}, support = 0.33, confidence = 1$$

$$\{x_8\} \to \{x_4\}, support = 0.33, confidence = 0.75$$

$$\{x_5\} \to \{x_7\}, support = 0.22, confidence = 1$$

$$\{x_5\} \to \{x_9\}, support = 0.22, confidence = 1$$

$$\{x_5\} \to \{x_7, x_9\}, support = 0.22, confidence = 1$$

$$\{x_5, x_9\} \to \{x_7\}, support = 0.22, confidence = 1$$

$$\{x_5, x_7\} \to \{x_9\}, support = 0.22, confidence = 1$$

$$\{x_9\} \to \{x_5\}, support = 0.22, confidence = 0.67$$

$$\{x_9\} \to \{x_7\}, support = 0.22, confidence = 0.67$$

$$\{x_9\} \to \{x_5, x_7\}, support = 0.22, confidence = 0.67$$

$$\{x_7, x_9\} \to \{x_5\}, support = 0.22, confidence = 1$$

$$\{x_6, x_7\} \to \emptyset, support = 0.22, confidence = 1$$

في هذه المجموعة النهائية من قواعد الاقتران، لا تخبرنا كل قاعدة اقتران في الشكل $\emptyset$ $\longrightarrow F$ عن الاقتران بين مجموعتي عناصر ولكن عن وجود مجموعة العناصر F في مجموعة البيانات، ومن ثم يمكن تجاهلها. تكشف قواعد الاقتران المتبقية عن الارتباط الوثيق لـ x_4 مع x_8 و x_5 مع x_7 و x_9 و x_6 مع x_7 ، الأمر الذي يتطابق مع تدفقات الإنتاج في الشكل ١٢-١. ومع ذلك، لا يتم إيجاد تدفقات الإنتاج من الآلة الأولى $M1$، والثانية $M2$ والثالثة $M3$ في مجموعات العناصر المتكررة ولا في المجموعة النهائية من قواعد الاقتران بسبب الطريقة التي يتم فيها أخذ عينات مجموعة البيانات من خلال النظر في جميع الأعطال

الآلية الأحادية. وحيث إن الآلة الأولى *M1*، والثانية *M2* والثالثة *M3* هي في بداية تدفقات الإنتاج ويتأثرن بأنفسهن فقط، فإن كل من x_1، x_2، x_3 تظهر بشكل أقل تكراراً في مجموعة البيانات مقارنةً بـ x_4 إلى x_9 ولنفس السبب، تكون قيمة الثقة *(confidence)* لقاعدة الاقتران $\{x_8\} \rightarrow \{x_4\}$ أعلى من تلك لقاعدة الاقتران $\{x_4\} \rightarrow \{x_8\}$.

يتم تطبيق اكتشاف قاعدة الاقتران على البيانات الرقمية. ولتطبيق اكتشاف قاعدة الاقتران، تحتاج البيانات الرقمية إلى أن يتم تحويلها إلى بيانات نوعية من خلال تعريف نطاقات قيم البيانات كما تمّ مناقشته في الجزء ٤-٣ من الفصل ٤ ومعاملة القيم في نفس النطاق باعتبارها من العنصر نفسه.

٣-١٢ البرمجيات والتطبيقات (Software and Applications):

يتم يدعم اكتشاف قاعدة الاقتران من خلال استخدام برنامج

Weka (*http://www.cs.waikato. ac.nz/ml/weka*).

والبرنامج

Statistica (*www.statistica.com*)

يمكن العثور على بعض تطبيقات قاعدة الاقتران في يي *(Ye, 2003, Chapter 2)*.

التمارين (Exercises):

١٢-١ انظر في سجلات البيانات الـ ١٦ في مجموعة البيانات الاختبارية لاكتشاف أعطال النظام في الجدول ٣-٢ باعتبارها ١٦ مجموعة من العناصر، من خلال أخذ x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 كتسع مشاكل جودة مختلفة وبقيمة 1 تشير إلى وجود مشكلة جودة معينة. أوجد جميع مجموعات العناصر المتكررة ذات الحد الأدنى للدعم *min-support=0.2*.

١٢-٢ استخدم مجموعات العناصر المتكررة من التمرين ١٢-١ لتوليد جميع قواعد الاقتران التي تحقق الحد الأدنى للدعم *min-support=0.2* والحد الأدنى للثقة *min-confidence=0.5*.

١٢-٣ كرِّر التمرين ١٢-١ لجميع سجلات البيانات البالغة ٢٥ من الجدول ١٢-١ والجدول ٣-٢ باعتبارها مجموعة البيانات.

١٢-٤ كرِّر التمرين ١٢-٢ لجميع سجلات البيانات البالغة ٢٥ من الجدول ١٢-١ والجدول ٣-٢ باعتبارها مجموعة البيانات.

١٢-٥ لتوضيح أن خوارزمية أبريوري (الأسبقية) تُعد فعّالة لمجموعة بيانات متناثرة، قم بإيجاد أو إنشاء مجموعة بيانات متناثرة بحيث يكون كل عنصر غير متكرر نسبياً في مجموعة البيانات، وقم بتطبيق خوارزمية أبريوري (الأسبقية) على مجموعة البيانات لاستخراج مجموعات عناصر متكررة وبقيمة مناسبة للحد الأدنى للدعم *min-support*.

١٢-٦ لتوضيح أن خوارزمية أبريوري (الأسبقية) تُعد أقل فعالية لمجموعة بيانات كثيفة، قم بإيجاد أو إنشاء مجموعة بيانات كثيفة بحيث يكون كل عنصر متكرراً نسبياً في سجلات بيانات مجموعة البيانات، وقم بتطبيق خوارزمية أبريوري (الأسبقية) على مجموعة البيانات لاستخراج مجموعات عناصر متكررة وبقيمة مناسبة للحد الأدنى للدعم *min-support*.

١٣- شبكة بـيـيز
Bayesian network

يتطلب مصنّف بييز (*Bayes classifier*) في الفصل ٣ أن تكون جميع متغيرات الخاصية مستقلة عن بعضها البعض. شبكة بييز (*Bayesian network*) في هذا الفصل تسمح بالاقتران (*association*) بين متغيرات الخاصية نفسها وبالاقتران بين متغيرات الخاصية ومتغيرات الهدف. تستخدم شبكة بييز اقتران المتغيرات لاستنتاج المعلومات عن أي متغير في شبكة بييز. في هذا الفصل، نستعرض البنية (*structure*) الخاصة بشبكة بييز ومعلومات الاحتمال الخاصة بالمتغيرات في شبكة بييز. ثم نقوم بوصف الاستدلال الاحتمالي (*probablitistic inference*) الذي يتم إجراؤه داخل شبكة بييز. وأخيراً، نستعرض طرق تعلُّم البنية ومعلومات الاحتمال الخاصة بشبكة بييز. وترد قائمة من حزم البرمجيات التي تدعم شبكة بييز. ويتم إعطاء بعض تطبيقات شبكة بييز مع مراجعها.

١٣-١ بُنية شبكة بييز والتوزيعات الاحتمالية للمتغيرات
(Structure of a Bayesian Network and Probability Distributions of Variables):

في الفصل ٣، يستخدم مصنّف بييز البسيط (*naïve Bayes classifier*) المعادلة ٣-٥ (كما سيتم توضيحها لاحقاً) لتصنيف قيمة متغير الهدف y على أساس افتراض أن متغيرات الخاصية، $x_i,...x_p$ تكون مستقلة عن بعضها البعض:

$$y_{MAP} \approx \arg\ max_{y \in Y}\ p(y) \prod_{i=1}^{p} P(x_i|y).$$

ومع ذلك، ففي كثير من التطبيقات، تقترن بعض متغيرات الخاصية بطريقة معينة. على سبيل المثال، في مجموعة بيانات اكتشاف أعطال النظام المبينة في الجدول ٣-١ والتي تمّ نسخها هنا في الجدول ١٣-١، تقترن x_1مع x_5 x_7 ، و x_9 كما هو موضح في الشكل ١-١، والتي تمّ نسخها هنا في الشكل ١٣-١، تكون الآلات الخامسة *M5*، والسابعة *M7*، والتاسعة *M9* على خط إنتاج وحدات المنتج التي يتم معالجتها في الآلة الأولى *M1*. ولكن الآلة الأولى المعطلة *M1* تتسبب في تراجع جودة وحدات المنتج بعد مرورها من الآلة الأولى *M1* حيث إن $x_1=1$ والذي بدوره يسبب أن تكون $x_5=1$ ثم$x_7=1$ ، وأخيراً $x_9=1$. وعلى الرغم من أن x_1 تؤثر

على x_5 x_7 ، و x_9 فإن كل من x_5 x_7 ، و x_9 لا تؤثر على x_1 ومن ثم، فإن اقتران السبب-التأثير (*cause-effect association*) لـ x_1 مع x_5 x_7 ، و x_9 يتجه في اتجاه واحد فقط. علاوةً على ذلك، لا تقترن x_1 مع المتغيرات الأخرى، x_2 x_3 x_4 x_6 وx_8

تحتوي شبكة بييز على عُقَد *(nodes)* لتمثيل المتغيرات (بما في ذلك متغيرات الخاصية - *attribute variables* - ومتغيرات الهدف - *attribute variables*) وروابط موجهة بين العُقَد لتمثيل الاقترانات الموجهة بين المتغيرات. وبفرض أن يكون لكل متغير مجموعة محدودة من الحالات أو القيم. يوجد رابط موجة من عقدة تمثل المتغير x_i إلى عقدة تمثل المتغير x_j إذا كانت x_i لها تأثير مباشر على x_j على سبيل المثال، x_i تسبب x_j أو يؤثر x_i على x_j بطريقة ما. في رابط موجه من x_i إلى x_j تكون x_i هي أب لـ x_j و x_j هي ابن لـ x_i من غير المسموح وجود دوائر موجهة (*directed cycles*)، على سبيل المثال $x_1 \rightarrow x_2 \rightarrow$ $x_3 \rightarrow$ x_1 في شبكة بييز. ومن ثم، فإن بنية شبكة بييز هي رسم بياني مفتوح وموجه .(*directed, acyclic graph*)

الجدول (١٣-١)

مجموعة البيانات التدريبية الخاصة باكتشاف أعطال نظام تصنيع

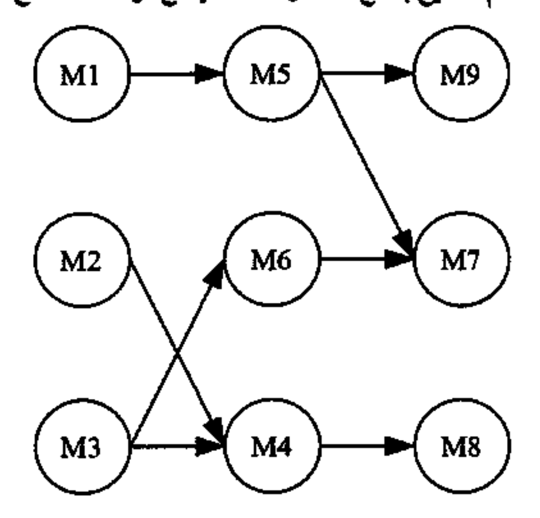
رقم الحالة - Instance (الآلة المعطلة - (Faulty Machine	Attribute Variables – متغيرات الخاصية Quality of Parts – جودة وحدات المنتج									متغير الهدف - Target Variable
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	عطل النظام (System Fault), y
1 (M1)	1	0	0	0	1	0	1	0	1	1
2(M2)	0	1	0	1	0	0	0	1	0	1
3(M3)	0	0	1	1	0	1	1	1	0	1
4(M4)	0	0	0	1	0	0	0	1	0	1
5(M5)	0	0	0	0	1	0	1	0	1	1
6(M6)	0	0	0	0	0	1	1	0	0	1
7(M7)	0	0	0	0	0	0	1	0	0	1
8(M8)	0	0	0	0	0	0	0	1	0	1
9(M9)	0	0	0	0	0	0	0	0	1	1
10(none)	0	0	0	0	0	0	0	0	0	0

وعادةً ما يتم استخدام مجال المعرفة (الذي تمّ جمع البيانات منه) لتحديد كيفية ارتباط المتغيرات. على سبيل المثال، تدفق إنتاج وحدات المنتج في الشكل ١٣-١ يمكن استخدامه لتحديد بنية شبكة بييز الموضحة في الشكل ١٣-٢ والتي تتضمن تسعة متغيرات خاصية لجودة وحدات المنتج في مراحل مختلفة من الإنتاج، x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 و x_9 ومتغير الهدف للإشارة لوجود أعطال بالنظام، y. في الشكل ١٣-٢، x_5 لديها أب واحد x_1 و x_6 لديها أب واحد x_3 و x_4 لديها أبوان x_2 و x_3 و x_9 لديها أب واحد x_5 و x_7 لديها أبوان x_5 و x_7 و x_8 لديها أب واحد x_4 و y لديها ثلاثة آباء x_7 و x_8 و x_9. بدلاً من رسم رابط موجه من كل من متغيرات الجودة التسعة x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 إلى متغير أعطال النظام y، فإن لدينا رابط موجه من كل من متغيرات الجودة الثلاثة، x_7 x_8 x_9 إلى متغير أعطال النظام y، نظراً لأن x_7 x_8 و x_9 في المرحلة الأخيرة من تدفق الإنتاج وتأخذ التأثير من x_1 x_2 x_3 x_4 x_5, x_6 على y.

إذا كان لدينا المتغير x وله الآباء z_k...، z_1 فإن شبكة بييز تستخدم التوزيع الاحتمالي المشروط لـ (*conditional probability distribution*) $p(x_1 | z_1, \ldots z_k)$ لقياس تأثير الآباء $z_1, \ldots z_k$ على الابن x على سبيل المثال، فإننا نفترض أن الجهاز المستخدم لفحص جودة وحدات المنتج في مجموعة بيانات اكتشاف أعطال نظام التصنيع لا يتم الاعتماد عليه 100 ٪، مما يؤدي إلى إنتاج بيانات غير يقينية *(data uncertainties)* وتوزيعات احتمالية مشروطة في الجداول من ١٣-٢ وحتى ١٣-١٠ للعُقَد التي لها أب (آباء) في الشكل ١٣-٢. على سبيل المثال، في الجدول ١٣-٢، $p(x_5=0 | x_1=1)=0.1$ و $p(x_5=1 | x_1=1)=0.9$ تعني أنه إذا كانت $x_1=1$ فإن احتمال $x_5=0$ هو 0.1، واحتمال أن $x_5=1$ هو 0.9، واحتمال وجود أي من هاتين القيمتين (0 أو 1) لـ x_5 هو $0.1+0.9=1$. يعود سبب عدم حصولنا على الاحتمالية 1 لـ $x_5=1$ إذا كانت $x_1=1$ إلى أن جهاز الفحص لـ x_1 لدية احتمال صغير للتعطل. وعلى الرغم من أن أجهزة الفحص تشير إلى أن $x_1=1$ إلا أن هناك احتمالاً صغيراً بأن x_1 يجب أن تكون صفراً. وبالإضافة إلى ذلك، فإن جهاز الفحص لـ x_5 لديه أيضاً احتمال صغير للتعطل، وهذا يعني أن جهاز الفحص ربما يشير إلى أن $x_5=0$ على الرغم من أن x_5 ينبغي أن تكون 1. احتمالات التعطل لأجهزة الفحص تنتج بيانات غير يقينية، ومن ثم يكون لدينا الاحتمالات المشروطة في الجداول من ١٣-٢ وحتى ١٣-١٠.

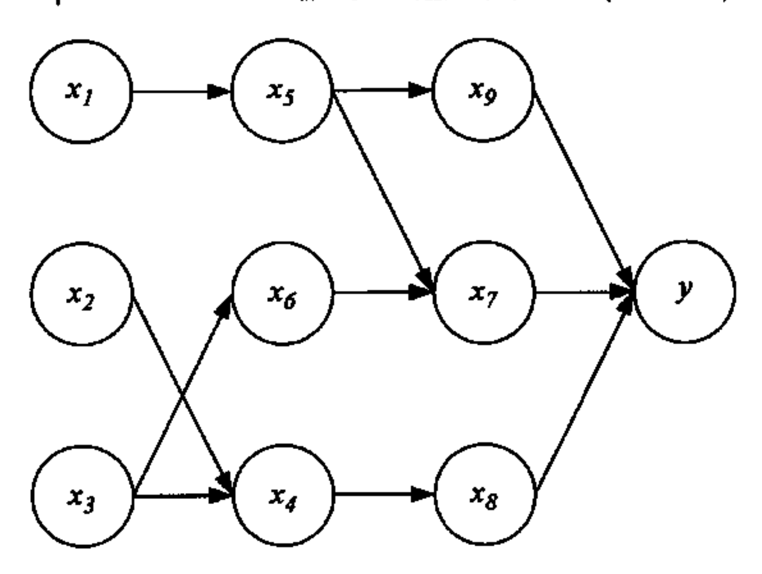
الشكل (١٣-١)

نظام تصنيع بتسع آلات وتدفقات إنتاج لوحدات المنتج



الشكل (١٣-٢)

البنية (structure) الخاصة بشبكة بييز لمجموعة بيانات اكتشاف أعطال نظام التصنيع



الجدول (١٣-٢)

$P(x_5| x_1)$

$x_1 = 1$	$x_1 = 0$	
$P(x_5 = 0 \| x_1 = 1) = 0.1$	$P(x_5 = 0 \| x_1 = 0) = 0.7$	$x_5 = 0$
$P(x_5 = 1 \| x_1 = 1) = 0.9$	$P(x_5 = 1 \| x_1 = 0) = 0.3$	$x_5 = 1$

الجدول (١٣-٣)

$P(x_6| x_3)$

$x_3 = 1$	$x_3 = 0$	
$P(x_6 = 0 \| x_3 = 1) = 0.1$	$P(x_6 = 0 \| x_3 = 0) = 0.7$	$x_6 = 0$
$P(x_6 = 1 \| x_3 = 1) = 0.9$	$P(x_6 = 1 \| x_3 = 0) = 0.3$	$x_6 = 1$

الجدول (١٣-٤)

$P(x_4| x_3, x_2)$

$x_2 = 0$		
$x_3 = 1$	$x_3 = 0$	
$P(x_4 = 0 \| x_2 = 0, x_3 = 1) = 0.1$	$P(x_4 = 0 \| x_2 = 0, x_3 = 0) = 0.7$	$x_4 = 0$
$P(x_4 = 1 \| x_2 = 0, x_3 = 1) = 0.9$	$P(x_4 = 1 \| x_2 = 0, x_3 = 0) = 0.3$	$x_4 = 1$
$x_2 = 1$		
$x_3 = 1$	$x_3 = 0$	
$P(x_4 = 0 \| x_2 = 1, x_3 = 1) = 0.1$	$P(x_4 = 0 \| x_2 = 1, x_3 = 0) = 0.1$	$x_4 = 0$
$P(x_4 = 1 \| x_2 = 1, x_3 = 1) = 0.9$	$P(x_4 = 1 \| x_2 = 1, x_3 = 0) = 0.9$	$x_4 = 1$

الجدول (١٣-٥)

$P(x_9 | x_5)$

$x_5 = 1$	$x_5 = 0$	
$P(x_9 = 0 \| x_5 = 1) = 0.1$	$P(x_9 = 0 \| x_5 = 0) = 0.7$	$x_9 = 0$
$P(x_9 = 1 \| x_5 = 1) = 0.9$	$P(x_9 = 1 \| x_5 = 0) = 0.3$	$x_9 = 1$

الجدول (١٣-٦)

$P(x_7 | x_5, x_6)$

$x_5 = 0$		
$x_6 = 1$	$x_6 = 0$	
$P(x_7 = 0 \| x_5 = 0, x_6 = 1) = 0.1$	$P(x_7 = 0 \| x_5 = 0, x_6 = 0) = 0.7$	$x_7 = 0$
$P(x_7 = 1 \| x_5 = 0, x_6 = 1) = 0.9$	$P(x_7 = 1 \| x_5 = 0, x_6 = 0) = 0.3$	$x_7 = 1$
$x_5 = 1$		
$x_6 = 1$	$x_6 = 0$	
$P(x_7 = 0 \| x_5 = 1, x_6 = 1) = 0.1$	$P(x_7 = 0 \| x_5 = 1, x_6 = 0) = 0.1$	$x_7 = 0$
$P(x_7 = 1 \| x_5 = 1, x_6 = 1) = 0.9$	$P(x_7 = 1 \| x_5 = 1, x_6 = 0) = 0.9$	$x_7 = 1$

الجدول (١٣-٧)

$P(x_8 | x_4)$

$x_4 = 1$	$x_4 = 0$	
$P(x_8 = 0 \| x_4 = 1) = 0.1$	$P(x_8 = 0 \| x_4 = 0) = 0.7$	$x_8 = 0$
$P(x_8 = 1 \| x_4 = 1) = 0.9$	$P(x_8 = 1 \| x_4 = 0) = 0.3$	$x_8 = 1$

الجدول (١٣-٨)

		$P(y\| x_9)$
$x_9 = 1$	$x_9 = 0$	
$P(y = 0\|x_9 = 1) = 0.1$	$P(y = 0\|x_9 = 0) = 0.9$	$y = 0$
$P(y = 1\|x_9 = 1) = 0.9$	$P(y = 1\|x_9 = 0) = 0.1$	$y = 1$

الجدول (١٣-٩)

		$P(y\| x_7)$
$x_7 = 1$	$x_7 = 0$	
$P(y = 0\|x_7 = 1) = 0.1$	$P(y = 0\|x_7 = 0) = 0.9$	$y = 0$
$P(y = 1\|x_7 = 1) = 0.9$	$P(y = 1\|x_7 = 0) = 0.1$	$y = 1$

الجدول (١٣-١٠)

		$P(y\| x_8)$
$x_8 = 1$	$x_8 = 0$	
$P(y = 0\|x_8 = 1) = 0.1$	$P(y = 0\|x_8 = 0) = 0.9$	$y = 0$
$P(y = 1\|x_8 = 1) = 0.9$	$P(y = 1\|x_8 = 0) = 0.1$	$y = 1$

بالنسبة لعقدة المتغير x في شبكة بييز التي لا يوجد لديها آباء، هناك حاجة للتوزيع الاحتمالي السابق (*prior probability distribution*) لـ x على سبيل المثال، في شبكة بييز في الشكل ١٣-٢، فإن x_1 x_2 ، و x_3 ،ليس لها آباء ويتم إعطاء التوزيعات الاحتمالية السابقة الخاصة بهم في الجداول من ١٣-١١ وحتى ١٣-١٣ على التوالي.

التوزيعات الاحتمالية السابقة الخاصة بالعُقَد التي ليس لها أب (آباء) والتوزيعات الاحتمالية المشروطة الخاصة بالعُقَد التي لها أب (آباء) تسمح بحساب التوزيع الاحتمالي المشترك (*joint probability distribution*) لجميع المتغيرات في شبكة بييز.

الجدول (١٣-١١)

	$P(x_1)$
$x_1 = 1$	$x_1 = 0$
$P(x_1 = 1) = 0.2$	$P(x_1 = 0) = 0.8$

الجدول (١٣-١٢)

	$P(x_2)$
$x_2 = 1$	$x_2 = 0$
$P(x_2 = 1) = 0.2$	$P(x_2 = 0) = 0.8$

الجدول (١٣-١٣)

	$P(x_3)$
$x_3 = 1$	$x_3 = 0$
$P(x_3 = 1) = 0.2$	$P(x_3 = 0) = 0.8$

على سبيل المثال، يتم حساب توزيع الاحتمال المشترك للمتغيرات الـ ١٠ في شبكة بييز في الشكل ١٣-٢ كما يلي:

$$P(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, y)$$

$$= P(y|x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)P(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)$$

$$= P(y|x_7, x_8, x_9)P(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)$$

$$= P(y|x_7, x_8, x_9)P(x_9|x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)P(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$$

$$= P(y|x_7, x_8, x_9)P(x_9|x_5)P(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$$

$$= P(y|x_7, x_8, x_9)P(x_9|x_5)P(x_7|x_1, x_2, x_3, x_4, x_5, x_6, x_8)P(x_1, x_2, x_3, x_4, x_5, x_6, x_8)$$

$$= P(y|x_7, x_8, x_9)P(x_9|x_5)P(x_7|x_5, x_6)P(x_1, x_2, x_3, x_4, x_5, x_6, x_8) = \cdots$$

$$= P(y|x_7, x_8, x_9)P(x_9|x_5)P(x_7|x_5, x_6)P(x_8|x_4)P(x_5|x_1)P(x_6|x_3)P(x_4|x_2, x_3)P(x_1, x_2, x_3)$$

$$= P(y|x_7, x_8, x_9)P(x_9|x_5)P(x_7|x_5, x_6)P(x_8|x_4)P(x_5|x_1)P(x_6|x_3)P(x_4|x_2, x_3)P(x_1)P(x_2)P(x_3)$$

في طريقة الحساب المذكورة أعلاه، نقوم باستخدام المعادلات التالية:

$$P(x_1, \dots, x_i|z_1, \dots, z_k, v_1, \dots, v_j) = P(x_1, \dots, x_i|z_1, \dots, z_k) \qquad (١٣-١)$$

$$P(x_1, \dots, x_i) = \prod_{j=1}^{i} P(x_i), \qquad (١٣-٢)$$

حيث إنه في المعادلة ١٣-١ لدينا $x_1, \dots, x_i$ مستقلة بشكل مشروط عن $v_1, \dots, v_j$ إذا علمنا قيم $z_1, \dots, z_k$ وفي المعادلة ١٣-٢ لدينا $x_1, \dots, x_i$ مستقلة عن بعضها البعض.

ومن ثم، فإن الاستقلال المشروط والاستقلال بين بعض المتغيرات يسمح لنا أن نعبر عن توزيع الاحتمال المشترك لجميع المتغيرات باستخدام توزيعات الاحتمال المشروط الخاص بالعُقَد التي لديها أب (آباء) وتوزيعات الاحتمال السابقة الخاصة بالعُقَد التي ليس لديها أب (آباء). وبعبارة أخرى، فإن شبكة بييز تعطي تمثيلاً مفككاً ومبسطاً لتوزيع الاحتمال المشترك.

توزيع الاحتمال المشترك لجميع المتغيرات يعطي الوصف الكامل لجميع المتغيرات ويسمح لنا بالإجابة عن أية أسئلة عن كل المتغيرات. على سبيل المثال، إذا كان لدينا توزيع الاحتمال المشترك لمتغيرين x وz، $P(x, z)$، وx تأخذ واحدة من القيم $a_1, \dots, a_i$ وz تأخذ واحدة من القيم $b_1, \dots, b_j$؛ يمكننا حساب الاحتمالات عن أي أسئلة عن هذين المتغيرين:

$$P(x) = \sum_{k=1}^{j} P(x, z = b_k) \qquad (١٣-٣)$$

$$P(z) = \sum_{k=1}^{i} P(x = a_k, z) \tag{٤-١٣}$$

$$P(x|z) = \frac{P(x,z)}{P(z)} \tag{٥-١٣}$$

$$P(z|x) = \frac{P(x,z)}{P(x)} \tag{٦-١٣}$$

في المعادلة ١٣-٣، نقوم بتهميش z من $P(x, z)$ للحصول على $P(x)$ في المعادلة ١٣-٤، نقوم بتهميش x من $P(x,z)$ للحصول على $P(z)$.

المثال (١٣-١):

إذا كان لدينا توزيع الاحتمال المشترك التالي $p(x,z)$:

$$P(x = 0, z = 0) = 0.2$$
$$P(x = 0, z = 1) = 0.4$$
$$P(x = 1, z = 0) = 0.3$$
$$P(x = 1, z = 1) = 0.1$$

والتي مجموعهم يساوي 1، احسب كل من $P(x)$ و$P(z)$ و$P(x|z)$ و$P(x|z)$:

$$P(x = 0) = P(x = 0, z = 0) + P(x = 0, z = 1) = 0.2 + 0.4 = 0.6$$

$$P(x = 1) = P(x = 1, z = 0) + P(x = 1, z = 1) = 0.3 + 0.1 = 0.4$$

$$P(z = 0) = P(x = 0, z = 0) + P(x = 1, z = 0) = 0.2 + 0.3 = 0.5$$

$$P(z=1)=P(x=0,z=1)+P(x=1,z=1)=0.4+0.1=0.5$$

$$P(x=0|z=0)=\frac{P(x=0,z=0)}{P(z=0)}=\frac{0.2}{0.5}=0.4$$

$$P(x=1|z=0)=\frac{P(x=1,z=0)}{P(z=0)}=\frac{0.3}{0.5}=0.6$$

$$P(x=0|z=1)=\frac{P(x=0,z=1)}{P(z=1)}=\frac{0.4}{0.5}=0.8$$

$$P(x=1|z=1)=\frac{P(x=1,z=1)}{P(z=1)}=\frac{0.1}{0.5}=0.2$$

$$P(z=0|x=0)=\frac{P(x=0,z=0)}{P(x=0)}=\frac{0.2}{0.6}=0.33$$

$$P(z=1|x=0)=\frac{P(x=0,z=1)}{P(x=0)}=\frac{0.4}{0.6}=0.67$$

$$P(z=0|x=1)=\frac{P(x=1,z=0)}{P(x=1)}=\frac{0.3}{0.4}=0.75$$

$$P(z=1|x=1)=\frac{P(x=1,z=1)}{P(x=1)}=\frac{0.1}{0.4}=0.25.$$

٢-١٣ الاستدلال الاحتمالي (Probabitistic Inference):

تمثل التوزيعات الاحتمالية المستنبطة من شبكة بييز معرفتنا السابقة عن مجال جميع المتغيرات. بعد الحصول على أدلة لقيم معينة لبعض المتغيرات (متغيرات الأدلة - *evidence variables*)، نريد أن نستخدم الاستدلال الاحتمالي لتحديد التوزيعات الاحتمالية اللاحقة (*posterior probability distribution*) الخاصة بالمتغيرات المستهدفة (متغيرات الاستعلام - *query varrable*). وهو ما يعني، أننا نريد أن نرى كيف

تتغير احتمالات القيم لمتغيرات الاستعلام بعد معرفة قيم معينة لمتغيرات الأدلة. على سبيل المثال، في شبكة بييز في الشكل ١٣-٢، نريد أن نعرف ما هو احتمال أن $y=1$، وما هو احتمال x_7 إذا كان لدينا الدليل المؤكد أن $x_9=1$ في بعض التطبيقات، متغيرات الدليل هي المتغيرات التي يمكن رصدها بسهولة، ومتغيرات الاستعلام هي المتغيرات التي لا يمكن رصدها. نعطي بعض الأمثلة على الاستدلال الاحتمالي.

المثال (١٣-٢):

بالنظر إلى شبكة بييز في الشكل ١٣-٢ والتوزيعات الاحتمالية في الجداول من ١٣-٢ إلى ١٣-١٣. إذا علمنا أنَّ $x_6=1$ ما هي احتمالات $x_4=1$ ، $x3=1$ ، و$x_2=1$ ؟

وبعبارة أخرى، ما هي $P(x_4 = 1|x_6 = 1)$ ،$P(x_3 = 1|x_6 = 1)$ و $P(x_2 = 1|x_6 = 1)$ ؟ لاحظ أن الشــرط المُعطى $x_6 = 1$ لا يعني أن $P(x_6 = 1) = 1$.

للحصول على $P(x_3 = 1|x_6 = 1)$، نحتاج الحصول على $P(x_3, x_6)$.

$P(x_6 , x_3) = P(x_6 | x_3)P(x_3)$

$x_3 = 1$	$x_3 = 0$	
$P(x_6 = 0, x_3 = 1) = P(x_6 = 0\|x_3 = 1)$ $P(x_3 = 1) = (0.1)(0.2) = 0.02$	$P(x_6 = 0, x_3 = 0) = P(x_6 = 0\|x_3 = 0)$ $P(x_3 = 0) = (0.7)(0.8) = 0.56$	$x_6 = 0$
$P(x_6 = 1, x_3 = 1) = P(x_6 = 1\|x_3 = 1)$ $P(x_3 = 1) = (0.9)(0.2) = 0.18$	$P(x_6 = 1, x_3 = 0) = P(x_6 = 1\|x_3 = 0)$ $P(x_3 = 0) = (0.3)(0.8) = 0.24$	$x_6 = 1$

من خلال تهميش x_3 خارج $P(x_6, x_3)$، نحصل على $P(x_6)$:

$$P(x_6 = 0) = P(x_6 = 0, x_3 = 0) + P(x_6 = 0, x_3 = 1) = 0.56 + 0.02 = 0.58$$

$$P(x_6 = 1) = P(x_6 = 1, x_3 = 0) + P(x_6 = 1, x_3 = 1) = 0.24 + 0.18 = 0.42.$$

$$P(x_3 = 1|x_6 = 1) = \frac{P(x_6 = 1|x_3 = 1)P(x_3 = 1)}{P(x_6 = 1)} = \frac{(0.9)(0.2)}{0.42} = 0.429$$

ومن ثم، فإن الدليل $x_6 = 1$ يغير الاحتمال $x_3 = 1$ من 0.2 إلى 0.429.

للحصول على $P(x_4 = 1|x_6 = 1)$، نحتاج الحصول على $P(x_4, x_6)$. وتقترن x_4 و x_6 من خلال x_3. وعلاوةً على ذلك، فإن الاقتران بين x_4 و x_3 يستلزم x_2. ومن ثم، نريد تهميش x_3 وx_2 خارج $P(x_4, x_3, x_2|x_6 = 1)$، حيث:

$$P(x_4, x_3, x_2|x_6 = 1) = P(x_4|x_3, x_2)P(x_3|x_6 = 1)P(x_2)$$

$$= P(x_4|x_3, x_2)\frac{P(x_6 = 1|x_3)P(x_3)}{P(x_6 = 1)}P(x_2).$$

على الرغم من أن $P(x_4|x_3, x_2)$، $P(x_6|x_3)$، $P(x_3)$ و$P(x_2)$ معطاه في الجداول ١٣-٣، ١٣-٤، ١٣-١٢و ١٣-١٣، على التوالي، نحتـــاج أن نحســـب $P(x_6)$. بالإضــافة إلى حســـاب $P(x_6)$، نحتاج أيضاً إلى حساب $P(x_4)$ لنتمكن من مقارنة $P(x_4 = 1|x_6 = 1)$ مع $P(x_4)$.

للحصول على $P(x_4)$ و$P(x_6)$، نقوم أولاً بحساب الاحتمالات المشتركة $P(x_4, x_3, x_2)$ و$P(x_6, x_3)$، ومن ثمّ نقوم بتهميش x_3 و x_2 خارج $P(x_4, x_3, x_2)$، وx_3 خارج $P(x_6, x_3)$ على النحو التالي:

$$P(x_4, x_3, x_2) = P(x_4|x_3, x_2)P(x_3)P(x_2)$$

$x_2 = 0$		
$x_3 = 1$	$x_3 = 0$	
$P(x_4 = 0, x_3 = 1, x_2 = 0) = P(x_4 = 0\|x_3 = 1, x_2 = 0)$ $P(x_3 = 1)P(x_2 = 0) = (0.1)(0.2)(0.8) = 0.016$	$P(x_4 = 0, x_3 = 0, x_2 = 0) = P(x_4 = 0\|x_3 = 0, x_2 = 0)$ $P(x_3 = 0)P(x_2 = 0) = (0.7)(0.8)(0.8) = 0.448$	$x_4 = 0$
$P(x_4 = 1, x_3 = 1, x_2 = 0) = P(x_4 = 1\|x_3 = 1, x_2 = 0)$ $P(x_3 = 1)P(x_2 = 0) = (0.9)(0.2)(0.8) = 0.144$	$P(x_4 = 1, x_3 = 0, x_2 = 0) = P(x_4 = 1\|x_3 = 0, x_2 = 0)$ $P(x_3 = 0)P(x_2 = 0) = (0.3)(0.8)(0.8) = 0.192$	$x_4 = 1$
$x_2 = 1$		
$x_3 = 1$	$x_3 = 0$	
$P(x_4 = 0, x_3 = 1, x_2 = 1) = P(x_4 = 0\|x_3 = 1, x_2 = 1)$ $P(x_3 = 1)P(x_2 = 1) = (0.1)(0.2)(0.2) = 0.004$	$P(x_4 = 0, x_3 = 0, x_2 = 1) = P(x_4 = 0\|x_3 = 0, x_2 = 1)$ $P(x_3 = 0)P(x_2 = 1) = (0.1)(0.8)(0.2) = 0.016$	$x_4 = 0$
$P(x_4 = 1, x_3 = 1, x_2 = 1) = P(x_4 = 1\|x_3 = 1, x_2 = 1)$ $P(x_3 = 1)P(x_2 = 1) = (0.9)(0.2)(0.2) = 0.036$	$P(x_4 = 1, x_3 = 0, x_2 = 1) = P(x_4 = 1\|x_3 = 0, x_2 = 1)$ $P(x_3 = 0)P(x_2 = 1) = (0.9)(0.8)(0.2) = 0.144$	$x_4 = 1$

وبتهميش x_3 و x_2 خارج $P(x_4, x_3, x_2)$، نحصل على $P(x_4)$:

$$\begin{aligned} P(x_4 = 0) &= P(x_4 = 0, x_3 = 0, x_2 = 0) + P(x_4 = 0, x_3 = 1, x_2 = 0) \\ &+ P(x_4 = 0, x_3 = 0, x_2 = 1) + P(x_4 = 0, x_3 = 1, x_2 = 1) \\ &= 0.448 + 0.016 + 0.016 + 0.004 = 0.484 \end{aligned}$$

$$\begin{aligned} P(x_4 = 1) &= P(x_4 = 1, x_3 = 0, x_2 = 0) + P(x_4 = 1, x_3 = 1, x_2 = 0) \\ &+ P(x_4 = 1, x_3 = 0, x_2 = 1) + P(x_4 = 1, x_3 = 1, x_2 = 1) \\ &= 0.192 + 0.144 + 0.144 + 0.036 = 0.516. \end{aligned}$$

والآن نستخدم $P(x_6)$ لحساب $P(x_4, x_3, x_2|x_6 = 1)$:

$$P(x_4, x_3, x_2|x_6 = 1) = P(x_4|x_3, x_2)P(x_3|x_6 = 1)P(x_2)$$

$$= P(x_4|x_3, x_2)\frac{P(x_6 = 1|x_3)P(x_3)}{P(x_6 = 1)}P(x_2):$$

	$x_2 = 0$	
	$x_3 = 0$	$x_3 = 1$
$x_4 = 0$	$P(x_4 = 0 \mid x_3 = 0, x_2 = 0)$ $\frac{P(x_6 = 1\mid x_3 = 0)P(x_3 = 0)}{P(x_6 = 1)}$ $P(x_2 = 0)$ $= (0.7)\frac{(0.3)(0.8)}{0.42}(0.8)$ $= 0.32$	$P(x_4 = 0 \mid x_3 = 1, x_2 = 0)$ $\frac{P(x_6 = 1\mid x_3 = 1)P(x_3 = 1)}{P(x_6 = 1)}$ $P(x_2 = 0)$ $= (0.1)\frac{(0.9)(0.2)}{0.42}(0.8)$ $= 0.034$
$x_4 = 1$	$P(x_4 = 1 \mid x_3 = 0, x_2 = 0)$ $\frac{P(x_6 = 1\mid x_3 = 0)P(x_3 = 0)}{P(x_6 = 1)}$ $P(x_2 = 0)$ $= (0.3)\frac{(0.3)(0.8)}{0.42}(0.8)$ $= 0.137$	$P(x_4 = 1 \mid x_3 = 1, x_2 = 0)$ $\frac{P(x_6 = 1\mid x_3 = 1)P(x_3 = 1)}{P(x_6 = 1)}$ $P(x_2 = 0)$ $= (0.9)\frac{(0.9)(0.2)}{0.42}(0.8)$ $= 0.309$

	$x_2 = 1$	
$x_3 = 1$	$x_3 = 0$	
$(x_4 = 0 \mid x_3 = 1, x_2 = 1)$ $\frac{P(x_6 = 1 \mid x_3 = 1)P(x_3 = 1)}{P(x_6 = 1)}$ $P(x_2 = 1)$ $= (0.1)\frac{(0.9)(0.2)}{0.42}(0.2)$ $= 0.009$	$P(x_4 = 0 \mid x_3 = 0, x_2 = 1)$ $\frac{P(x_6 = 1 \mid x_3 = 0)P(x_3 = 0)}{P(x_6 = 1)}$ $P(x_2 = 1)$ $= (0.1)\frac{(0.3)(0.8)}{0.42}(0.2)$ $= 0.011$	$x_4 = 0$
$(x_4 = 1 \mid x_3 = 1, x_2 = 1)$ $\frac{P(x_6 = 1 \mid x_3 = 1)P(x_3 = 1)}{P(x_6 = 1)}$ $P(x_2 = 1)$ $= (0.9)\frac{(0.9)(0.2)}{0.42}(0.2)$ $= 0.077$	$P(x_4 = 1 \mid x_3 = 0, x_2 = 1)$ $\frac{P(x_6 = 1 \mid x_3 = 0)P(x_3 = 0)}{P(x_6 = 1)}$ $P(x_2 = 1)$ $= (0.9)\frac{(0.3)(0.8)}{0.42}(0.2)$ $= 0.103$	$x_4 = 1$

نحصل على $P(x_4 = 1|x_6 = 1)$ من خلال تهميش x_3 و x_2 خارج $P(x_4, x_3, x_2|x_6 = 1)$:

$$
\begin{aligned}
&P(x_4 = 1|x_6 = 1) \\
&= P(x_4 = 1, x_3 = 0, x_2 = 0|x_6 = 1) \\
&+ P(x_4 = 1, x_3 = 1, x_2 = 0|x_6 = 1) \\
&+ P(x_4 = 1, x_3 = 0, x_2 = 1|x_6 = 1) \\
&+ P(x_4 = 1, x_3 = 1, x_2 = 1|x_6 = 1) \\
&= 0.137 + 0.309 + 0.103 + 0.077 = 0.626.
\end{aligned}
$$

بمقارنة $P(x_4 = 1) = 0.516$ التي قمنا بحسابها سابقاً، فإنّ الدليل $x_6 = 1$ يغير الاحتمال $x_4 = 1$ إلى 0.626.

نحصل على $P(x_2 = 1|x_6 = 1)$ من خلال تهميش x_4 و x_3 خارج $P(x_4, x_3, x_2|x_6 = 1)$:

$$
\begin{aligned}
&P(x_2 = 1|x_6 = 1)\\
&= P(x_4 = 0, x_3 = 0, x_2 = 1|x_6 = 1)\\
&+ P(x_4 = 1, x_3 = 0, x_2 = 1|x_6 = 1)\\
&+ P(x_4 = 0, x_3 = 1, x_2 = 1|x_6 = 1)\\
&+ P(x_4 = 1, x_3 = 1, x_2 = 1|x_6 = 1)\\
&= 0.011 + 0.103 + 0.009 + 0.077 = 0.2.
\end{aligned}
$$

الدليل على أن $x_6 = 1$ لا يغير الاحتمــال أن $x_2 = 1$ من الاحتمــال السابق 0.2، لأن x_6 يتأثر بـ x_3 فقط. الدليل على أن $x_6 = 1$ يجلب الحاجة إلى تحديث الاحتمال اللاحق لـ x_3 والذي بدوره يجلب الحاجة إلى تحديث الاحتمال اللاحق لـ x_4، لأن x_3 يؤثر على x_4

وبشكل عام، قمنا بإجراء الاستدلال الاحتمالي عن متغير استعلام (*quesry variable*) عن طريق الحصول أولاً على التوزيع الاحتمالي المشترك الذي يحتوي على متغير الاستعلام، ومن ثمّ تهميش المتغيرات غير المستعلم عنها (*non query variables*) خارج التوزيع الاحتمالي المشترك للحصول على احتمال متغير الاستعلام. بغض النظر عما إذا تم الحصول على دليل جديد عن قيمة معينة لمتغير، فإن التوزيع الاحتمالي المشروط لا يتغير لكل عقدة لها أب (آباء)، احتمال حدوث الابن (*child*) علماً بحدوث الأب (*parent*) أو الآباء $P(child \mid parent(s))$ والمعطاة في شبكة بييز، ومع ذلك، فإن جميع الاحتمالات الأخرى، بما في ذلك الاحتمالات المشروطة *p(parent child)* واحتمالات المتغيرات الأخرى غير المتغير الدليل، قد تتغير، اعتماداً على ما إذا كانت تلك الاحتمالات قد تأثرت بالمتغير الدليل أم لا.

كل الاحتمالات التي تتأثر بمتغير الدليل تحتاج إلى تحديث، وينبغي أن تُستخدَم الاحتمالات المحدَّثة للاستدلال الاحتمالي عندما يتم الحصول على أدلة جديدة. على سبيل المثال، إذا واصلنا من المثال ١٣-٢ وحصلنا على دليل جديد $x_4 = 1$ بعد تحديث الاحتمالات للدليل $x_6 = 1$ في المثال ١٣-٢، فإن جميع الاحتمالات التي تمّ تحديثها من المثال ١٣-٢ ينبغي أن تُستخدَم لإجراء الاستدلال الاحتمالي للدليل الجديد $x_4 = 1$ على سبيل المثال الاستدلال الاحتمالي لتحديد $P(x_3 = 1|x_4 = 1)$ و $P(x_2 = 1|x_4 = 1)$.

المثال (١٣-٣):

بالاستمرارية مع جميع الاحتمالات اللاحقة المُحدَّثة للدليل $x_6 = 1$ من المثال ١٣-٢، نحصل الآن على دليل جديد: $x_4 = 1$ ما الاحتمالات اللاحقة لـ $x_2 = 1$ و $x_3 = 1$؟ وبعبارة أخرى، عند البدء بجميع الاحتمالات التي تمّ تحديثها من المثال ١٣-٢، ما هي $P(x_3 = 1|x_4 = 1)$ و $P(x_2 = 1|x_4 = 1)$ ؟

يتم استعراض الاستدلال الاحتمالي لاحقاً:

$$P(x_3, x_2|x_4 = 1) = \frac{P(x_4 = 1|x_3, x_2)P(x_3|x_6 = 1)P(x_2|x_6 = 1)}{P(x_4 = 1|x_6 = 1)} = \frac{(0.9)(0.2)}{0.42} = 0.429$$

$$P(x_3 = 0, x_2 = 0|x_4 = 1) = \frac{P(x_4 = 1|x_3 = 0, x_2 = 0)P(x_3 = 0|x_6 = 1)P(x_2 = 0|x_6 = 1)}{P(x_4 = 1|x_6 = 1)}$$

$$= \frac{(0.3)(1 - 0.429)(1 - 0.2)}{(0.626)} = 0.219$$

$$P(x_3 = 0, x_2 = 1|x_4 = 1) = \frac{P(x_4 = 1|x_3 = 0, x_2 = 1)P(x_3 = 0|x_6 = 1)P(x_2 = 1|x_6 = 1)}{P(x_4 = 1|x_6 = 1)}$$

$$= \frac{(0.9)(1 - 0.429)(0.2)}{(0.626)} = 0.164$$

$$P(x_3 = 1, x_2 = 0|x_4 = 1) = \frac{P(x_4 = 1|x_3 = 1, x_2 = 0)P(x_3 = 1|x_6 = 1)P(x_2 = 0|x_6 = 1)}{P(x_4 = 1|x_6 = 1)}$$

$$= \frac{(0.9)(0.429)(1 - 0.2)}{(0.626)} = 0.494$$

$$P(x_3 = 1, x_2 = 1|x_4 = 1) = \frac{P(x_4 = 1|x_3 = 1, x_2 = 1)P(x_3 = 1|x_6 = 1)P(x_2 = 1|x_6 = 1)}{P(x_4 = 1|x_6 = 1)}$$

$$= \frac{(0.9)(0.429)(0.2)}{(0.626)} = 0.123$$

نحصل على $P(x_3 = 1|x_4 = 1)$ من خلال تهميش x_2 خارج $P(x_3, x_2|x_4 = 1)$:

$$P(x_3 = 1|x_4 = 1) = P(x_3 = 1, x_2 = 0|x_4 = 1) + P(x_3 = 1, x_2 = 1|x_4 = 1)$$

$$= 0.494 + 0.123 = 0.617$$

بما أن x_3 تؤثر على كل من x_6 و x_4 نرفع احتمال أن $x_3 = 1$ من 0.2 إلى 0.429، عندما يكون لدينا الدليل $x_6 = 1$ ثم نرفع احتمال أن $x_3 = 1$ مرةً أخرى من 0.429 إلى 0.617 عندما يكون لدينا الدليل $x_4 = 1$.

وبذلك نحصل على $P(x_2 = 1|x_4 = 1)$ من خلال تهميش x_3 خارج $P(x_3, x_2|x_4 = 1)$:

$$P(x_2 = 1|x_4 = 1) = P(x_3 = 0, x_2 = 1|x_4 = 1) + P(x_3 = 1, x_2 = 1|x_4 = 1)$$

$$= 0.164 + 0.123 = 0.287.$$

بما أن x_2 تؤثر على x_4 ولكن لا تؤثر على x_6، يبقى احتمال $x_2 = 1$ هو نفسه عند 0.2 عندما يكون لدينا الدليل على $x_6 = 1$ ثم نرفع احتمال أن $x_2 = 1$ من 0.2 إلى 0.287، عندما يكون لدينا الدليل على $x_4 = 1$ وهي ليست زيادة كبيرة لأن $x_3 = 1$ قد تنتج أيضا الدليل على $x_4 = 1$.

تحتاج الخوارزميات التي تُستخدَم لعمل الاستدلال الاحتمالي للبحث عن مسار من المتغير الدليل إلى متغير الاستعلام، وتحديث واستنتاج الاحتمالات على طول المسار، كما فعلنا ذلك يدوياً في الأمثلة ١٣-٢ و١٣-٣. ويتطلب البحث والاستدلال الاحتمالي القيام بكم كبير من الحسابات، كما رأينا في الأمثلة ١٣-٢ و١٣-٣. ومن ثم، لا بد من تطوير خوارزمية حاسوبية فعّالة لإجراء الاستدلال الاحتمالي في شبكة بييز، على سبيل المثال تلك الموجودة في *HUGIN* (*www.hugin.com*) وهي حزمة برمجية لشبكة بييز.

٣-١٣ تعلُّم شبكة بييز (Learning of a Bayesian Network):

إن تعلُّم البنية الخاصة بشبكة بييز والاحتمالات المشروطة والاحتمالات السابقة في شبكة بييز من بيانات استكشافية هو موضوع قيد البحث بشكل واسع. وبشكل عام، نود أن نقوم بتركيب بنية شبكة بييز على أساس مجال المعرفة قيد البحث. ولكن، عندما لا يكون لدينا معرفة كافية عن المجال المبحوث والمستهدف، ولكن لدينا فقط بيانات عن بعض المتغيرات المرصودة في المجال، فنحن بحاجة للكشف عن الاقترانات بين المتغيرات باستخدام تقنيات استكشاف البيانات، مثل: قواعد الاقتران (*association rules*) الموجودة في الفصل ١٢، والأساليب الإحصائية، مثل: إجراء اختبارات على استقلالية المتغيرات.

عندما تكون جميع المتغيرات في شبكة بييز قابلة للرصد للحصول على سجلات بيانات للمتغيرات، فإنه يمكن تقدير جداول الاحتمالية المشروطة للعُقَد التي لها أب (آباء) والاحتمالات السابقة للعُقَد دون أب (آباء)، باستخدام الصيغ التالية كما هو الحال في المعادلات ٣-٦ و٣-٧:

$$P(x = a) = \frac{N_{x=a}}{N} \qquad (٧-١٣)$$

$$P(x = a|z = b) = \frac{N_{x=a\&z=b}}{N_{z=b}}, \qquad (٨-١٣)$$

حيث إن:

N هو عدد سجلات البيانات في مجموعة البيانات.

$N_{x=a}$ هو عدد سجلات البيانات مع $x = a$

$N_{z=b}$ هو عدد نقاط البيانات مع $z = b$

$N_{x=a\&z=b}$ هو عدد سجلات البيانات مع $x = a$ و $z = b$

وقد قام راسيل وآخرون (*Russell et al., 1995*) بتطوير طريقة الصعود المتدرج (*gradient ascent method*)، والتي تشبه طريقة الهبوط المتدرج (*gradient*

(*decent method* للشبكة العصبية الصناعية، لتعلم المُدخَل (*entry*) في جدول الاحتمال المشــروط في شبــكة بييــز، عنــدما لا يمكن تعلـم المُــدخَل من الـبيـانات الاستكشافية أو التدريبية. فليــكن $w_{ij} = P(x_i|z_j)$ عبارة عن مدخل في جدول الاحتمال المشروط للعقدة x التي تأخذ القيمة رقم i والتي لها أب (أباء) z والتي تأخذ القيمة رقم j في شبكة بييز. ولتكن h تشير إلى فرضية عن قيمة w_{ij}. إذا كان لدينا مجموعة بيانات استكشافية، نريد أن توجد فرضية الإمكان الأكبر (*maximum likelihood hypothesis*) h التي تعظِّم قيمة $P(D|h)$.

$$h = \arg\ max_h\ P(D|h) = \arg\ max_h\ ln\, P(D|h)\ .$$

يتم تنفيذ الصعود المتدرج التالي لتحديث w_{ij}:

$$w_{ij}(t+1) = w_{ij}(t+1) + \alpha \frac{\partial ln\, P(D|h)}{\partial w_{ij}}, \qquad (١٣-٩)$$

حيث α هو معدل التعلم. بترميز $P(D|h)$ إلى $P_h(D)$ واستخدام $\partial lnf(x)/\partial x = [1/f(x)][\partial f(x)/\partial x]$، يكون لدينا:

$$\frac{\partial ln\, P(D|h)}{\partial w_{ij}} = \frac{\partial ln\, P_h(D)}{\partial w_{ij}} = \frac{\partial ln \prod_{d\in D} P_h(d)}{\partial w_{ij}}$$

$$= \sum\nolimits_{d\in D} \frac{1}{P_h(d)} \frac{\partial P_h(d)}{\partial w_{ij}} = \sum\nolimits_{d\in D} \frac{1}{P_h(d)} \frac{\partial \sum_{i'j'} P_h(d|x_{i'}, z_{j'}) P_h(x_{i'}, z_{j'})}{\partial w_{ij}}$$

$$= \sum\nolimits_{d\in D} \frac{1}{P_h(d)} \frac{\partial \sum_{i'j'} P_h(d|x_{i'}, z_{j'}) P_h(x_{i'}|z_{j'})\, P_h(z_{j'})}{\partial w_{ij}}$$

$$= \sum\nolimits_{d\in D} \frac{1}{P_h(d)} \frac{\partial \sum_{i'j'} P_h(d|x_{i'}, z_{j'})\ w_{i'j'}\ P_h(z_{j'})}{\partial w_{ij}}$$

$$= \sum_{d \in D} \frac{1}{P_h(d)} P_h(d|x_i, z_j)\, P_h(z_j) = \sum_{d \in D} \frac{1}{P_h(d)} \frac{P_h(x_i, z_j|d) P_h(d)}{P_h(x_i, z_j)} P_h(z_j)$$

$$= \sum_{d \in D} \frac{P_h(x_i, z_j|d)}{P_h(x_i, z_j)} P_h(z_j) = \sum_{d \in D} \frac{P_h(x_i, z_j|d)}{P_h(x_i|z_j)} = \sum_{d \in D} \frac{P_h(x_i, z_j|d)}{w_{ij}}.$$

(١٠-١٣)

بإدخال المعادلة ١٣-١٠ في ١٣-٩، نحصل على:

$$w_{ij}(t+1) = w_{ij}(t+1) + \alpha \frac{\partial ln\, P(D|h)}{\partial w_{ij}} = w_{ij}(t+1) + \alpha \sum_{d \in D} \frac{P_h(x_i, z_j|d)}{w_{ij}(t)}, \qquad (11\text{-}13)$$

حيث $P_h(x_i, z_j|d)$ يمكن الحصول عليها باستخدام الاستدلال الاحتمالي الموضح في الجزء ١٣-٢. بعد استخدام المعادلة ١٣-١١ لتحديث w_{ij}، نحتاج إلى أن نتأكد من أن:

$$\sum_i w_{ij}(t+1) = 1 \qquad (12\text{-}13)$$

عن طريق إجراء التطبيع:

$$w_{ij}(t+1) = \frac{w_{ij}(t+1)}{\sum_i w_{ij}(t+1)}. \qquad (13\text{-}13)$$

١٣-٤ البرمجيات والتطبيقات (Software and Applications):

خادم بييز *(Bayes server)* (*www.bayesserver.com*) وهيوقن *(HUGIN)* (*www.hugin.com*) هما حزمتان برمجيتان تدعمان شبكة بييز. يمكن العثور على بعض التطبيقات الخاصة بشبكة بييز في مجال المعلومات الحيوية (*bioinformatics*)، وبعض المجالات الأخرى في ديفيز (*Davis, 2003*)، دييز وآخرون (*Diez et al., 1997*)، وجيانغ كوبر (*Jiang and Cooper, 2010*)، وبوريت وآخرون (*Pourret at al., 2008*).

التمارين (Exercises):

١١-١ بالنظر في شبكة بييز في الشكل ١٣-٢، والتوزيعات الاحتمالية في الجداول من ١٣-٢ إلى ١٣-١٣. وإذا كان لدينا $x_6 = 1$ ما هو احتمال أن $x_7 = 1$؟، وبعبارة أخرى، ما هو $P(x_1 = 1|x_6 = 1)$؟

١٣-١ بالنظر في شبكة بييز في الشكل ١٣-٢، والتوزيعات الاحتمالية في الجداول من ١٣-٢ إلى ١٣-١٣. وإذا كان لدينا $x_6 = 1$ ما هو احتمال أن $x_7 = 1$؟، وبعبارة أخرى، ما هو $P(x_1 = 1|x_6 = 1)$؟

١٣-٢ بالاستمرارية مع جميع الاحتمالات اللاحقة المحدثة للدليل $x6 = 1$ من المثال ١٣-٢ والمثال ١٣-١، نحصل الآن على دليل جديد $x4 = 1$ ما الاحتمال اللاحق $x7 = 1$؟ وبعبارة أخرى، ما هو $P(x1 =1| x4 =1)$؟

١٣-٣ كرّر التمرين ١٣-١ لتحديد $P(x1 = 1| x6 = 1)$.

١٣-٤ كرّر التمرين ١٣-٢ لتحديد $P(x1 = 1| x4 = 1)$.

١٣-٥ كرّر التمرين ١٣-١ لتحديد $P(y = 1| x6 = 1)$.

١٣-٦ كرّر التمرين ١٣-٢ لتحديد $P(y = 1| x4 = 1)$.

الجزء الرابع
خوارزميات استكشاف أنماط اختزال البيانات
Algorithms for Mining Data Reduction Patterns

١٤- تحليل المكونات الرئيسية
Principal Component Analysis

تحليل المكونات الرئيسية (PCA) هي تقنية إحصائية لتمثيل البيانات العالية الأبعاد في فضاء منخفض الأبعاد. وعادةً ما يتم استخدام تحليل المكونات الرئيسية (PCA) لاختزال أبعاد البيانات، بحيث يمكن تصوير أو تحليل البيانات في فضاء منخفض الأبعاد. على سبيل المثال، قد نستخدم تحليل المكونات الرئيسية (PCA) لتمثيل سجلات بيانات لها ١٠٠ متغير من متغيرات الخاصية بسجلات بيانات لها متغيران أو ثلاثة من المتغيرات. في هذا الفصل، يتم أولاً مراجعة إحصاءات المتغيرات المتعددة (*multivariate statistics*)، وجبر المصفوفات (*algebra matrix*) لوضع ومعرفة الأساس الرياضي لتحليل المكونات الرئيسية (PCA). وبعد ذلك، سيتم وصف وتوضيح تحليل المكونات الرئيسية (PCA). وترد قائمة بحزم البرمجيات التي تدعم تحليل المكونات الرئيسية (PCA). ويتم إعطاء بعض التطبيقات الخاصة بتحليل المكونات الرئيسية (PCA) مع مراجعها.

١-١٤ مراجعة لإحصاءات المتغيرات المتعددة
(Review of Multivariate Statistics):

إذا كان x_i عبارة عن متغير عشوائي متصل أو كمّي بقيم مستمرة وبدالة كثافة احتمال $f_i(x_i)$، فإن كلاً من المتوسط (*mean*)، u_i، والتباين (*variance*) σ_i^2 للمتغير العشوائي، يتم تعريفهما على النحو التالي:

$$u_i = E(x_i) = \int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i \qquad (١-١٤)$$

$$\sigma_i^2 = \int_{-\infty}^{\infty} (x_i - u_i)^2 f_i(x_i)\, dx_i. \qquad (٢-١٤)$$

إذا كان x_i عبارة عن متغير عشوائي غير متصل أو نوعي (*discrete random variable*) وبقيم غير متصلة ودالة احتمال $P(x_i)$،

$$u_i = E(x_i) = \sum_{\substack{all\ values\\ of\ x_i}} x_i P(x_i) \qquad (٣-١٤)$$

$$\sigma_i^2 = \sum_{\substack{all\ values\\ of\ x_i}} (x_i - u_i)^2 P(x_i). \qquad (٤-١٤)$$

إذا كان كل من x_i وx_j عبارة عن متغيرين عشوائيين متصلين أو كميين وبدالة كثافة احتمال مشتركة $f_{ij}(x_i, x_j)$، فإنه يتم تعريف التغاير أو التباين المشترك (*Covariance*) للمتغيرين العشوائيين ، x_i وx_j على النحو التالي:

$$\sigma_{ij} = E(x_i - \mu_i)(x_j - \mu_j) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - u_i)(x_j - u_j) f_{ij}(x_i, x_j)\, dx_i\, dx_j. \qquad (٥-١٤)$$

إذا كان، x_i وx_j عبارة عن متغيرين عشوائيين غير متصلين أو نوعيين وبدالة كثافة احتمال مشتركة $P(x_i, x_j)$،

$$\sigma_{ij} = E(x_i - \mu_i)(x_j - \mu_j) = \sum_{\substack{all\ values\\ of\ x_i}} \sum_{\substack{all\ values\\ of\ x_j}} (x_i - \mu_i)(x_j - \mu_j) P(x_i, x_j). \qquad (٦-١٤)$$

ومعامل الارتباط (*correlation coefficient*) هو:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i}\sqrt{\sigma_j}}. \qquad (٧-١٤)$$

بالنسبة لمتجه (*vector*) المتغيرات العشوائية، $x = (x_1, x_2, \ldots, x_p)$ فإن المتجه المتوسط (*mean vector*) هو:

$$E(x) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}, \qquad (٨-١٤)$$

ومصفوفة التباين- التغاير (*Variance -Covariance*) هي:

$$\Sigma = E(x-\boldsymbol{\mu})(x-\boldsymbol{\mu})' = E\left(\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_p-\mu_p \end{bmatrix}\begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 & \cdots & x_p-\mu_p \end{bmatrix}\right)$$

$$= E\begin{pmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1)(x_p-\mu_p) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 & \cdots & (x_2-\mu_2)(x_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (x_p-\mu_p)(x_1-\mu_1) & (x_p-\mu_p)(x_2-\mu_2) & \cdots & (x_p-\mu_p)^2 \end{pmatrix}$$

$$= E\begin{pmatrix} E(x_1-\mu_1)^2 & E(x_1-\mu_1)(x_2-\mu_2) & \cdots & E(x_1-\mu_1)(x_p-\mu_p) \\ E(x_2-\mu_2)(x_1-\mu_1) & E(x_2-\mu_2)^2 & \cdots & E(x_2-\mu_2)(x_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(x_p-\mu_p)(x_1-\mu_1) & E(x_p-\mu_p)(x_2-\mu_2) & \cdots & E(x_p-\mu_p)^2 \end{pmatrix}$$

$$= \begin{bmatrix} \sigma_1 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p \end{bmatrix}. \qquad (١٤-٩)$$

المثال ١٤-١:

احسب المتجه المتوسط، ومصفوفة التباين - التغاير لاثنين من المتغيرات في الجدول ١٤-١. مجموعة البيانات في الجدول ١٤- ١ هي جزء من مجموعة البيانات الخاصة بنظام التصنيع في الجدول ١- ٤، وتحتوي على متغيري خاصية، x_7 و x_8 لتسع حالات من الأعطال الآلية الأحادية. ويبين الجدول ١٤- ٢ الاحتمالات المشتركة والهامشية لهذين المتغيرين.

المتوسط والتباين لـ x_7 هما:

$$u_7 = E(x_7) = \sum_{\substack{all\ values\\ of\ x_7}} x_7 P(x_7) = 0 \times \frac{4}{9} + 1 \times \frac{5}{9} = \frac{5}{9}$$

$$\sigma_7^2 = \sum_{\substack{all\ values\\ of\ x_7}} (x_7 - u_7)^2 P(x_7) = \left(0 - \frac{5}{9}\right)^2 \times \frac{4}{9} + \left(1 - \frac{5}{9}\right)^2 \times \frac{5}{9}$$

$$= 0.2469.$$

الجدول (١٤-١)

مجموعة البيانات الخاصة بالكشف عن الأعطال بنظام التصنيع مع متغيرين للجودة

x_8	x_7	رقم الحالة - Instance (الآلة المعطلة – Faulty Machine)
0	1	1 (M1)
1	0	2(M2)
1	1	3(M3)
1	0	4(M4)
0	1	5(M5)
0	1	6(M6)
0	1	7(M7)
1	0	8(M8)
0	0	9(M9)

الجدول (١٤-٢)

الاحتمالات المشتركة والهامشية لمتغيري الجودة

$P(x_7)$	x_8		$P(x_7, x_8)$
	1	0	x_7
$\frac{1}{9}+\frac{3}{9}=\frac{4}{9}$	$\frac{3}{9}$	$\frac{1}{9}$	0
$\frac{4}{9}+\frac{1}{9}=\frac{5}{9}$	$\frac{1}{9}$	$\frac{4}{9}$	1
1	$\frac{3}{9}+\frac{1}{9}=\frac{4}{9}$	$\frac{1}{9}+\frac{4}{9}=\frac{5}{9}$	$P(x_8)$

المتوسط والتباين لـ x_8 هما:

$$u_8 = E(x_8) = \sum_{\substack{all\ values \\ of\ x_8}} x_8 P(x_8) = 0 \times \frac{5}{9} + 1 \times \frac{4}{9} = \frac{4}{9}$$

$$\sigma_8^2 = \sum_{\substack{all\ values\\ of\ x_8}} (x_8 - u_8)^2 P(x_8) = \left(0 - \frac{4}{9}\right)^2 \times \frac{5}{9} + \left(1 - \frac{4}{9}\right)^2 \times \frac{4}{9}$$
$$= 0.2469.$$

التغاير (التباين المشترك) لكل من لـ x_7 و x_8 هو:

$$\sigma_{78} = \sum_{\substack{all\ values\\ of\ x_7}} \sum_{\substack{all\ values\\ of\ x_8}} (x_7 - \mu_7)(x_8 - \mu_8) P(x_7, x_8)$$

$$= \left(0 - \frac{5}{9}\right)\left(0 - \frac{4}{9}\right) \times \frac{1}{9} + \left(0 - \frac{5}{9}\right)\left(0 - \frac{4}{9}\right) \times \frac{3}{9} + \left(1 - \frac{5}{9}\right)\left(0 - \frac{4}{9}\right) \times \frac{4}{9}$$
$$+ \left(1 - \frac{5}{9}\right)\left(1 - \frac{4}{9}\right) \times \frac{1}{9} = -0.1358.$$

المتجه المتوسط $x = (x_7, x_8)$ هو:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \frac{5}{9} \\ \frac{4}{9} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{77} & \sigma_{78} \\ \sigma_{87} & \sigma_{88} \end{bmatrix} = \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}$$

٢-١٤ مراجعة جبر المصفوفات (Review of Matrix Algebra)

إذا كان لدينا متجه بعدد p من المتغيرات:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, \quad x' = [x_1 \quad x_2 \quad \cdots \quad x_p], \qquad (١٠-١٤)$$

تكون $x_1, x_2, \ldots, x_p$ غير مستقلة خطياً إذا وُجِـد مجموعة من الثوابت، $c_1, c_2, \ldots, c_p$، كلها لا تساوي الصفر، والتي تجعل المعادلة التالية صحيحة:

$$c_1x_1 + c_2x_2 + \cdots + c_px_p = 0. \qquad (١٤\text{-}١١)$$

بالمثل، فإنَّ $x_1, x_2, \ldots, x_p$ تعد مستقلة خطياً إذا وُجِـد مجموعة واحدة فقط من الثوابت، $c_1 = c_2 = \ldots = c_p = 0$، والتي تجعل المعادلة التالية صحيحة:

$$c_1x_1 + c_2x_2 + \cdots + c_px_p = 0. \qquad (١٤\text{-}١٢)$$

يتم حساب طول المتجه، x على النحو التالي:

$$L_x = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2} = \sqrt{x'x}. \qquad (١٤\text{-}١٣)$$

يوضح الشكل ١٤-١ متجهاً ثنائي الأبعاد، $x' = (x_1, x_2)$ ويُظهِر حساب طول المتجه. ويبين الشكل ١٤-٢ الزاوية θ بين متجهين، $x' = (x_1, x_2)$ و $y' = (y_1, y_2)$، والتي يتم حسابها على النحو التالي:

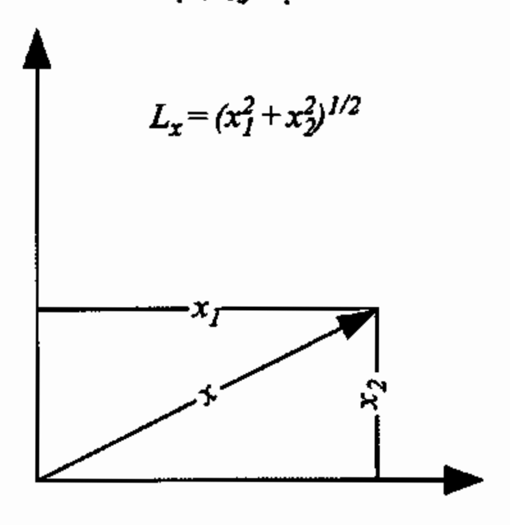
$$\cos(\theta_1) = \frac{x_1}{L_x} \qquad (١٤\text{-}١٤)$$

$$\sin(\theta_1) = \frac{x_2}{L_x} \qquad (١٤\text{-}١٥)$$

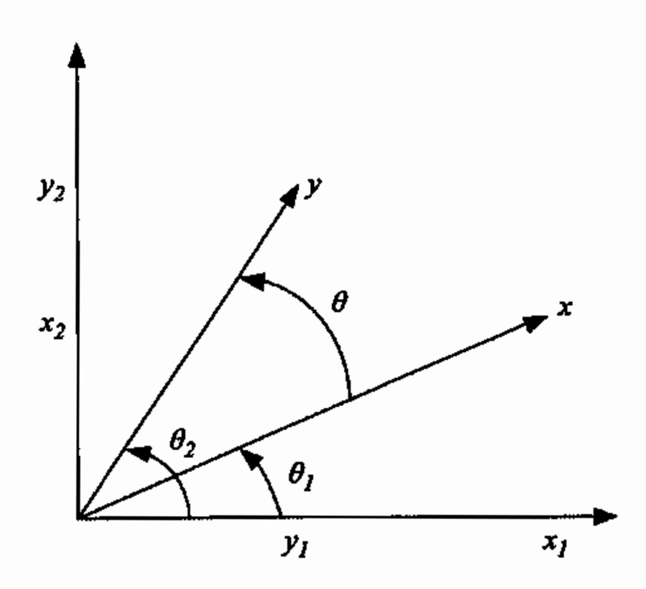
$$\cos(\theta_2) = \frac{y_1}{L_y} \qquad (١٤\text{-}١٦)$$

$$\sin(\theta_2) = \frac{y_2}{L_y} \qquad (١٤\text{-}١٧)$$

الشكل (١٤-١)
حساب طول المتجه



الشكل (١٤-٢)
حساب الزاوية بين متجهين



$$\cos(\theta) = \cos(\theta_2 - \theta_1)$$
$$= \cos(\theta_2)\cos(\theta_1)$$
$$+ \sin(\theta_2)\sin(\theta_1)$$

$$= \left(\frac{y_1}{L_y}\right)\left(\frac{x_1}{L_x}\right) + \left(\frac{y_2}{L_y}\right)\left(\frac{x_2}{L_x}\right) = \frac{x_1y_1 + x_2y_2}{L_xL_y} = \frac{x'y}{L_xL_y}. \quad (١٤-١٨)$$

وبناءً على عملية حساب الزاوية بين المتجهين، x' وy'، يكون المتجهان متعامدين، وهو ما يعني أن، $\theta = 90^o$ أو 270^o، أو $\cos(\theta) = 0$، فقط إذا كان $x'y = 0$.

وتكــون المصفــوفة المربعــة، $p \times p$، A متناظرة (*symmetric*) إذا كانت $A = A'$، وهو ما يعني أن، $a_{ij} = a_{ji}$، لكــل $i = 1, ..., p$، و $j = 1, ..., p$. والمصفوفة المحايدة (*Identity matrix*) تكون بالشكل التالي:

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

ويكون لدينا:

$$AI = \text{IA} = A. \quad (١٤-١٩)$$

ويُرمز لمعكوس المصفوفة A (*inverse of the matrix*) بالرمز على A^{-1}، ويكون لدينا:

$$AA^{-1} = A^{-1}A = I. \quad (١٤-٢٠)$$

ويوجد معكوس المصفوفة A إذا كانت أعمدة المصفوفة A والتي عددها k، ($a_1, a_2, \ldots, a_p$) مستقلةٌ خطياً.

ليكن $|A|$ يشير إلى مُحدد (*determinant*) المصفوفة A المربعة $p \times p$. يتم حساب $|A|$على النحو التالي:

$$|A| = a_{11} \quad if\ p = 1 \tag{٢١-١٤}$$

$$|A| = \sum_{j=1}^{p} a_{1j}|A_{1j}|\,(-1)^{1+j} = \sum_{j=1}^{p} a_{ij}|A_{ij}|\,(-1)^{i+j} \quad if p > 1, \tag{٢٢-١٤}$$

حيث إن:

A_{1j}	هي المصفوفة $(p-1) \times (p-1)$ التي تمّ الحصول عليها عن طريق إزالة الصف الأول والعمود j^{th} من A
A_{ij}	هي المصفوفة $(p-1) \times (p-1)$ التي تمّ الحصول عليها عن طريق إزالة الصف i^{th} والعمود j^{th} من A

ولمصفوفة مربعة ٢ × ٢:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

فإن محدد المصفوفة A هو:

$$\begin{aligned} |A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} &= \sum_{j=1}^{2} a_{1j}|A_{1j}|\,(-1)^{1+j} \\ &= a_{11}|A_{11}|(-1)^{1+1} + a_{12}|A_{12}|(-1)^{1+2} = a_{11}a_{22} - a_{12}a_{21}. \end{aligned} \tag{٢٣-١٤}$$

وبالنسبة للمصفوفة المحايدة I

$$|I| = 1. \tag{٢٤-١٤}$$

ويوضح التالي عملية حساب محدد المصفوفة A باستخدام مصفوفة التباين - التغاير لـ x_7 و x_8 من الجدول ١٤-١:

$$A = \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}$$
$$= 0.2469 \times 0.2469 - (-0.1358)(-0.1358)$$
$$= 0.0425.$$

لتكن A مصفوفة مربعة $p \times p$ وI المصفوفة المحايدة $p \times p$، فإن القيم $\lambda_1, \ldots, \lambda_p$تُسمى القيم الذاتية (وتُسمى أحياناً بقيم أيجن أو الجذور الكامنة) (***eigenvalues***) للمصفوفة A إذا كانت تُحقق المعادلة التالية:

$$|A - \lambda I| = 0. \qquad (١٤\text{-}٢٥)$$

المثال ١٤-٢:

احسب القيم الذاتية للمصفوفة A التالية، والتي يتم الحصول عليها من المثال ١٤-١:

$$A = \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}$$

$$|A - \lambda I| = \left| \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right|$$
$$= \begin{vmatrix} 0.2469 - \lambda & -0.1358 \\ -0.1358 & 0.2469 - \lambda \end{vmatrix} = 0.$$
$$(0.2469 - \lambda)(0.2469 - \lambda) - 0.0184 = 0$$

$$\lambda^2 - 0.4938\lambda + 0.0426 = 0$$

$$\lambda_1 = 0.3824 \qquad \lambda_2 = 0.1115.$$

لتكن A مصفوفة مربعة $p \times p$، وλ هي القيمة الذاتية لـ A. المتجه x يكون المتجه الذاتي (*eigenvector*) لــ A والمرتبط بالقيمة الذاتية λ إذا كان x هو متجه غير صفري ويحقق المعادلة التالية:

$$Ax = \lambda x. \qquad (١٤\text{-}٢٦)$$

يتم حساب المتجه الذاتي المطبّع (*normalized eigenvector*) بوحدة طول ، e، على النحو التالي:

$$e = \frac{x}{\sqrt{x'x}}. \qquad (١٤\text{-}٢٧)$$

المثال ١٤-٣

احسب المتجهات الذاتية المرتبطة بالقيم الذاتية في المثال ١٤-٢. يتم حساب المتجهات الذاتية المرتبطة بالقيم الذاتية $\lambda_1 = 0.3824$ و$\lambda_2 = 0.1115$ للمصفوفة المربعة التالية A في المثال ١٤-٢:

$$A = \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}$$

$$Ax = \lambda_1 x$$

$$\begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.3824 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{cases} 0.2469x_1 - 0.1358x_2 = 0.3824x_1 \\ -0.1358x_1 + 0.2469x_2 = 0.3824x_2 \end{cases}$$

$$\begin{cases} 0.1355x_1 + 0.1358x_2 = 0 \\ 0.1358x_1 + 0.1355x_2 = 0. \end{cases}$$

وبما أنَّ المعادلتين متطابقتان، فإنه هنالك العديد من الحلول. بوضع $x_1 = 1$ و $x_2 = -1$ يكون لدينا:

$$x = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad e = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$$

$$Ax = \lambda_2 x$$

$$\begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.1115 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{cases} 0.2469x_1 - 0.1358x_2 = 0.1115x_1 \\ -0.1358x_1 + 0.2469x_2 = 0.1115x_2 \end{cases}$$

$$\begin{cases} 0.1354x_1 + 0.1358x_2 = 0 \\ 0.1358x_1 + 0.1354x_2 = 0. \end{cases}$$

المعادلتان المذكورتان أعلاه متطابقتان، ومن ثم يكون لهما العديد من الحلول. بوضع $x_1 = 1$، و $x_2 = 1$ يصبح لدينا:

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad e = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

في هذا المثال، يتم اختيار المتجهين الذاتيين المرتبطين بالقيمتين الذاتيتين بحيث يكون المتجهان الذاتيان متعامدين.

لتكن A مصفوفة متطابقة $p \times p$، و (λ_i, e_i)، بحيث $i = 1, \ldots, p$، وتمثل p عدد p من أزواج القيم الذاتية والمتجهات الذاتية لـ A، بحيث أنَّ e_i، $i = 1, \ldots, p$ ، يتم اختياره

ليكون متعامداً بشكل متبادل. يُعطي التحلل الطيفي (*spectral decomposition*) للمصفوفة A بالمعادلة التالية:

$$A = \sum_{i=1}^{p} \lambda_i e_i e_i^{'}. \qquad (١٤\text{-}٢٨)$$

المثال ١٤-٤:

احسب التحلل الطيفي للمصفوفة في الأمثلة ١٤-٢ و١٤-٣.

يتم توضيح التحلل الطيفي للمصفوفة المتطابقة التالية في الأمثلة ١٤-٢ و١٤-٣ كما يلي:

$$A = \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}$$

$$\lambda_1 = 0.3824 \quad \lambda_2 = 0.1115$$

$$e_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$$

$$e_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$\begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}$$

$$= 0.3824 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

$$+ 0.1115 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} 0.1912 & -0.1912 \\ -0.1912 & 0.1912 \end{bmatrix} + \begin{bmatrix} 0.0558 & 0.0558 \\ 0.0558 & 0.0558 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1912 & -0.1912 \\ -0.1912 & 0.1912 \end{bmatrix} + \begin{bmatrix} 0.0558 & 0.0558 \\ 0.0558 & 0.0558 \end{bmatrix}$$

$$A = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2'.$$

وتُسمَى المصفوفة A المتطابقة $p \times p$، بالمصفوفة المحددة الموجبة (*positive definite matrix*) إذا حققت التالي لأي متجه غير صفري = $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

$$x'Ax > 0.$$

المصفوفة A المتطابقة $p \times p$ هي مصفوفة محددة موجبة إذاً وإذا كانت فقط كل قيمة ذاتية لـــ A أكبر من أو تساوي الصفر (*Johnson and Wichern, 1998*). على سبيل المثال، المصفوفة التالية A، 2×2 ، هي مصفوفة محددة موجبة بقيمتين ذاتيتين موجبتين:

$$A = \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}$$

$$\lambda_1 = 0.3824 \quad \lambda_2 = 0.1115$$

لتكن A مصفوفة محددة موجبة $p \times p$ بقيم ذاتية مرتبة كالتالي $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ وبقيم ذاتية مطبعة مرتبطة، $e_1, e_2, \ldots, e_p$، والتي تكون متعامدة. الشكل التربيعي، $(x'Ax)/(x'x)$، يتم تعظيمـــه إلى القيمة λ_1 عندما $x = e_1$ وهذا الشكل التربيعي يتم تصغيره إلى القيمة λ_p عندما $x = e_p$، (*Johnson and Wichern, 1998*).
وهو ما يعني، أن لدينا ما يلي:

$$\max_{x\neq 0} \frac{x'Ax}{x'x} = \lambda_1 \quad \text{attained by } x = e_1$$

أو

$$e_1'Ae_1 = e_1'\left(\sum_{i=1}^{p} \lambda_i e_i e_i'\right)e_1 = \lambda_1 = \max_{x\neq 0} \frac{x'Ax}{x'x} \qquad (٢٩\text{-}١٤)$$

$$\min_{x\neq 0} \frac{x'Ax}{x'x} = \lambda_p \quad \text{attained by } x = e_p$$

أو

$$e_p'Ae_p = e_p'\left(\sum_{i=1}^{p} \lambda_i e_i e_i'\right)e_p = \lambda_p = \min_{x\neq 0} \frac{x'Ax}{x'x} \qquad (٣٠\text{-}١٤)$$

و

$$\max_{x\perp e_1,\ldots,e_i} \frac{x'Ax}{x'x} = \lambda_{i+1} \quad \text{attained by } x = e_{i+1}, \quad i = 1, \ldots, p-1 \qquad (٣١\text{-}١٤)$$

١٤-٣ تحليل المكونات الرئيسة (Principal Component Analysis):

يوضح تحليلُ المكونات الرئيسية مصفوفةَ التباين- التغاير للمتغيرات. إذا كان لدينا متجه متغيرات $x'= [x_1,...,x_p]$ مع مصفوفة التباين- التغاير Σ، فيما يلي يمثل تركيباً خطياً لهذه المتغيرات:

$$y_i = a_i'x = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p. \quad (١٤-٣٢)$$

يمكن حساب التباين والتغاير لـ y_i على النحو التالي:

$$\text{var}(y_i) = a_i'\Sigma a_i \quad (١٤-٣٣)$$

$$\text{cov}(y_i, y_j) = a_i'\Sigma a_j. \quad (١٤-٣٤)$$

يتم اختيار المكونات الرئيسية $y'=[y_1, y_2, ..., y_p]$ لتكوّن تركيبات خطية لـ x' والتي تحقق ما يلي:

$$y_1 = a_1'x = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p,$$

$a'_1a_1=1$ ، يتم اختيار a_1 لتعظيم قيمة $var(y_1)$ (١٤-٣٥)

$$y_2 = a_2'x = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p,$$

$a'_2a_2=1, cov(y_2,y_1)=0$ ، يتم اختيار a_2 لتعظيم قيمة $var(y_2)$

$$\vdots$$

$$y_i = a_i'x = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p.$$

$a'_ia_i=1, cov(y_i,y_j)=0$ ، لكل $j<i$ ، يتم اختيار a_i لتعظيم قيمة $var(y_i)$

لتكن $(\lambda_i, e_i),\ i = 1, \ldots, p$، قيماً ذاتية ومتجهات ذاتية متعامدة لـ Σ، $e_i'e_i = 1$ و $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. بوضع $a_1 = e_1, \ldots, a_p = e_p$، يكون لدينا:

$$
\begin{aligned}
y_i &= e_i'x \quad i = 1, \ldots, p \\
e_i'e_i &= 1 \\
\text{var}(y_i) &= e_i'\Sigma e_i = \lambda_i \\
\text{cov}(y_i, y_j) &= e_i'\Sigma e_j = 0 \quad \text{for } j < i.
\end{aligned}
\tag{٣٦-١٤}
$$

بناءً على المعادلات من ١٤-٢٩ إلى ١٤-٣١، فإنّ y_i، $i=1,\ldots,p$ ، والمعدّلة بالمعادلة ١٤-٣٦، تحقق متطلبات المكونات الرئيسية في المعادلة ١٤-٣٥. بالتالي، يتم تحديد المكونات الرئيسية باستخدام المعادلة ١٤-٣٦.

لنجعل x_p ،... x_1 لها التباينات σ_p ،... σ_1 على التوالي. يكون مجموع التباينات x_p ،... x_1 مساوياً لمجموع تباينات y_p ،... y_1 (*Johnson and Wichern, 1998*):

$$
\sum_{i=1}^{p} \text{var}(x_i) = \sigma_1 + \cdots + \sigma_p = \sum_{i=1}^{p} \text{var}(y_i) = \lambda_1 + \cdots + \lambda_p. \tag{٣٧-١٤}
$$

مثال ١٤-٥:

قم بتحديد المكونات الرئيسية للمتغيرين في المثال ١٤-١. للمتغيرين $x' = [x_7, x_8]$ في الجدول ١٤-١ والمثال ١٤-١، تكون مصفوفة التباين- التغاير Σ على النحو التالي:

$$
\Sigma = \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix},
$$

وباستخدام القيم الذاتية والمتجهات الذاتية المحددة في الأمثلة ١٤-٢ و١٤-٣:

$$\lambda_1 = 0.3824 \qquad \lambda_2 = 0.1115$$

$$e_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$$

$$e_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}.$$

تكون المكونات الرئيسية:

$$y_1 = e_1'x = \frac{1}{\sqrt{2}}x_7 - \frac{1}{\sqrt{2}}x_8$$

$$y_2 = e_2'x = \frac{1}{\sqrt{2}}x_7 + \frac{1}{\sqrt{2}}x_8.$$

وتكون التباينات لـ y_1 وy_2 :

$$\begin{aligned}\mathrm{var}(y_1) &= \mathrm{var}\left(\frac{1}{\sqrt{2}}x_7 - \frac{1}{\sqrt{2}}x_8\right)\\ &= \left(\frac{1}{\sqrt{2}}\right)^2 \mathrm{var}(x_7) + \left(\frac{-1}{\sqrt{2}}\right)^2 \mathrm{var}(x_8) + 2\left(\frac{1}{\sqrt{2}}\right)\left(\frac{-1}{\sqrt{2}}\right)\mathrm{cov}(x_7, x_8)\\ &= \frac{1}{2}(0.2469) + \frac{1}{2}(0.2469) - (-0.1358) = 0.3827\\ &= \lambda_1\end{aligned}$$

$$\text{var}(y_2) = \text{var}\left(\frac{1}{\sqrt{2}}x_7 + \frac{1}{\sqrt{2}}x_8\right)$$

$$= \left(\frac{1}{\sqrt{2}}\right)^2 \text{var}(x_7) + \left(\frac{1}{\sqrt{2}}\right)^2 \text{var}(x_8) + 2\left(\frac{1}{\sqrt{2}}\right)\left(\frac{1}{\sqrt{2}}\right)\text{cov}(x_7, x_8)$$

$$= \frac{1}{2}(0.2469) + \frac{1}{2}(0.2469) + (-0.1358) = 0.1111$$

$$= \lambda_2$$

ويكون لدينا أيضاً:

$$\text{var}(x_7) + \text{var}(x_8) = 0.2469 + 0.2469 = \text{var}(y_1) + \text{var}(y_2) = 0.3827 + 0.1111.$$

وتكون نسبة مجموع التباينات المحتسبة في المكون الرئيسي الأول y_1 هي $0.3824/0.4939=0.7742$ أو 77٪. وحيث إن معظم مجموع التباينات في $x'[x_7\ x_8] =$ تمّ احتسابها بواسطة y_1، قد نستخدم y_1 ليحل محل وليمثل بالأساس المتغيرين $x_7\ x_8$ دون فقدان الكثير من التباينات. وهذا هو أساس تطبيق *PCA* لاختزال أبعاد البيانات باستخدام عدد قليل من المكونات الرئيسية لتمثيل عدد كبير من المتغيرات في البيانات الأصلية وفي الوقت نفسه يتم تمثيل الكثير من التباينات في البيانات. وباستخدام عدد قليل من المكونات الرئيسية لتمثيل البيانات، يمكن زيادة تصورنا للبيانات في فضاء أحادي، ثنائي، أو ثلاثي الأبعاد من المكونات الرئيسية لرصد أنماط البيانات، أو يمكن التنقيب أو البحث عنها أو تحليلها للكشف عن أنماط بيانات للمكونات الرئيسية. لاحظ أن المعنى الرياضي لكل مكون رئيسي كتركيب خطي لمتغير البيانات الأصلية ليس بالضرورة أن يكون له تفسيرٌ ذو مغزى في مجال المشكلة المبحوثة أو المستهدفة. يعطي يي (*Ye, 1997, 1998*) بعض الأمثلة لتفسير البيانات التي لا يتم تمثيلها في مجال المشكلة الأصلية.

١٤-٤ البرمجيات والتطبيقات (Software and Applications):

يتم دعم استخدام *PCA* من قبل العديد من حزم البرمجيات الإحصائية، بما في ذلك *SAS* (*www.sas.com*)، *SPSS* (*www.spss.com*)، و*STATISTICA* (*www.statistica.com*). ويتم إعطاء بعض تطبيقات *PCA* في المجالات الصناعية في يي (*Ye, 2003, Chapter 8*).

التمارين (Exercises):

١-١ قم بتحديد المكونات الرئيسية x_1 ...، x_9 في الجدول ٨-١ وتحديد المكونات الرئيسية التي يمكن استخدامها لتمثل ٩٠٪ من مجموع تباينات البيانات.

٢-١ قم بتحديد المكونات الرئيسية لـ x_1 و x_2 في الجدول ٣-٢.

٣-١ كرّر التمرين ١٤-٢ لـ x_1 ...، x_9 وحدد المكونات الرئيسية التي يمكن استخدامها لتمثل ٩٠٪ من مجموع تباينات البيانات.

١٥- القياس المتعدد الأبعاد
Multidimensional Scaling - MDS

يهدف القياس المتعدد الأبعاد (Multidimensional Scaling-MDS) إلى تمثيل البيانات عالية الأبعاد في فضاء منخفض الأبعاد بحيث يمكن تصور البيانات، وتحليلها، وتفسيرها في فضاء منخفض الأبعاد للكشف عن أنماط بيانات مفيدة. يصف هذا الفصل القياس المتعدد الأبعاد (MDS)، وحزم البرمجيات التي تدعمه، وبعض تطبيقاته مع المراجع المستخدمة.

١٥-١ خوارزمية القياس المتعدد الأبعاد (Algorithm of MDS):

ليكن مُعطًى لنا عدد n من عناصر البيانات في فضاء بعدد p من الأبعاد، $x_i=(x_{i1},\ldots,x_{ip})$، حيث أنَّ: $i=1,\ldots,n$، وبمقياس للاختلاف أو عدم التشابه (*dissimilarity*) δ_{ij} لكل زوج (x_i و x_j) من عناصر البيانات التي عددها n، وترتيب هذه الاختلافات من الزوج الأقل تشابهاً إلى الزوج الأكثر تشابهاً:

$$\delta_{i1j1} \le \delta_{i2j2} \le \cdots \le \delta_{iMjM}, \qquad (١٥\text{-}١)$$

حيث ترمز M إلى العدد الإجمالي لأزواج البيانات المختلفة، و، $M=n(n-1)/2$، لعدد n من عناصر البيانات وينبغي للقياس المتعدد الأبعاد (*MDS*) (*Young and Hamer, 1987*) إيجاد إحداثيات عناصر البيانات n في فضـــاء p من الأبعاد، $z_i=(z_{i1},\ldots,x_{iq}),i=1,\ldots,n$، وتكون q أصغر بكثير من p، مع المحافظة على اختلاف عناصر البيانات n الواردة في المعادلة ١٥-١. يكون القياس المتعدد الأبعاد (*MDS*) غير متري (*nonmetric*) إذا تمّ الحفاظ على ترتيب الاختلاف في المعادلة ١٥-١. ويذهب القياس المتعدد الأبعاد المتري (*metric*) إلى أبعد من ذلك ليحافظ على مقدار الاختلاف. يشرح هذا الفصل القياس المتعدد الأبعاد غير المتري.

يعرض الجدول ١٥-١ خطوات خوارزمية القياس المتعدد الأبعاد (*MDS*) لإيجاد إحداثيات عناصر البيانات n في فضاء بعدد q من الأبعاد، مع الحفاظ على اختلاف سجلات البيانات n الواردة في المعادلة ١٥-١. في الخطوة ١ من خوارزمية (*MDS*)، يتم توليد التهيئة الأولى لإحداثيات سجلات البيانات n في فضاء q من الأبعاد باستخدام قيم عشوائية بحيث لا يكون لسجلي بيانات القيم نفسها.

في الخطوة ٢ من خوارزمية (*MDS*)، يتم استخدام ما يلي لتطبيع، $x_i = (x_{i1}, \ldots, x_{iq})$ ، حيث إن $i = 1, \ldots, n$:

$$\text{normalized } x_{ij} = \frac{x_{ij}}{\sqrt{x_{i1}^2 + x_{iq}^2}} \quad (١٥\text{-}٢)$$

في الخطوة ٣ من خوارزمية (*MDS*)، يتم استخدام التالي لحساب ما يُسمى بجهد التهيئة (*stress of configuration*) الذي يقيس مدى جودة محافظة التهيئة على اختلاف سجلات البيانات n الواردة في المعادلة ١٥-١ (*Kruskal, 1964a,b*):

$$S = \sqrt{\frac{\sum_{ij}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{ij} d_{ij}^2}} \quad (١٥\text{-}٣)$$

حيث إن d_{ij} يقيس الاختلاف لـ x_i و x_j باستخدام إحداثياتها في فضاء بعدد q من الأبعاد، وتعطي القيمة $\hat{d}_{ij}$ الاختلاف المنشود لـ x_i و x_j الذي يحافظ على ترتيب الاختلاف لـ $\delta_{ij}s$ في المعادلة ١٥-١ بحيث يكون:

$$\hat{d}_{ij} < \hat{d}_{i'j'} \qquad \text{if } \hat{\delta}_{ij} < \hat{\delta}_{i'j'}. \quad (١٥\text{-}٤)$$

لاحظ أن هناك عدد $n\,(n\text{-}1)\,/2$ زوج مختلف من i و j في المعادلات ١٥-٣ و١٥-٤.

الجدول (١٥-١)

خوارزمية القياس المتعدد الأبعاد (*MDS*) - (إنجليزي وعربي)

Step	Description
1	Generate an initial configuration for the coordinates of n data points in the q-dimensional space, $(x_{11}, \ldots, x_{1q}, \ldots, x_{n1}, \ldots, x_{nq})$, such that no two points are the same
2	Normalize $x_i = (x_{i1}, \ldots, x_{iq})$, $i = 1, \ldots, n$, such that the vector for each data point has the unit length using Equation 15.2
3	Compute S as the stress of the configuration using Equation 15.3
4	REPEAT UNTIL a stopping criterion based on S is satisfied
5	Update the configuration using the gradient decent method and Equations 15.14 through 15.18
6	Normalize $x_i = (x_{i1}, \ldots, x_{iq})$, $i = 1, \ldots, n$, in the configuration using Equation 15.2
7	Compute S of the updated configuration using Equation 15.3

الخطوة	الوصف
١	قم بتوليد تهيئة أولية لإحداثيات سجلات البيانات n في فضاء q من الأبعاد $(x_{11}, \ldots, x_{1q}, \ldots, x_{n1} \ldots, x_{nq})$ بحيث لا يكون لسجلي بيانات القيم نفسها.
٢	قم بتطبيع، $x_i = (x_{i1}, \ldots, x_{iq})$ حيث أنَّ: $i=1, \ldots, n$، بحيث يكون لمتجه كل سجل بيانات نفس طول الوحدة باستخدام المعادلة ١٥-٢.
٣	قم بحساب s كقيمة لجهد التهيئة (*configuration Stress*) باستخدام المعادلة ١٥-٣.
٤	كرِّر (*REPEAT*) حتى (*UNTIL*) يتحقق شرط التوقف المبني على أساس قيمة S.
٥	حدث التهيئة (*configuration*) باستخدام طريقة الهبوط المتدرج والمعادلات من ١٥-١٤ إلى ١٥-١٨.
٦	قم بتطبيع، $x_i = (x_{i1}, \ldots, x_{iq})$ حيث إنَّ: $i=1, \ldots, n$، باستخدام المعادلة ١٥-٢.
٧	قم بحساب s للتهيئة المُحدَّثة باستخدام ١٥-٣.

ويمكن استخدام المسافة الإقليدية *(Euclidean distance)* الواردة في المعادلة ١٥-٥، أو مسافة مينكوسكي r المترية (*Minkowski r-metric distance*) الأكثر عموميةً في المعادلة ١٥-٦، أو يمكن استخدام بعض مقاييس الاختلاف الأخرى لحساب d_{ij}:

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left(d_{ik} - d_{jk}\right)^2} \qquad (٥\text{-}١٥)$$

$$d_{ij} = \left[\sum_{k=1}^{q} \left(d_{ik} - d_{jk}\right)^r\right]^{\frac{1}{r}}. \qquad (٦\text{-}١٥)$$

يتم التنبؤ بـقيم $\hat{d}_{ij}s$ من قيم $\delta_{ij}s$ باستخدام خوارزمية الانحدار الرئيسية الموضحة (*monotone regression algorithm)*) في (*Kruskal, 1964a,b*) لإعطاء:

$$\hat{d}_{i1j1} \leq \hat{d}_{i2j2} \leq \cdots \leq \hat{d}_{iMjM}, \qquad (٧\text{-}١٥)$$

وبالرجوع للمعادلة المعطاة في ١٥-١:

$$\delta_{i1j1} \leq \delta_{i2j2} \leq \cdots \leq \delta_{iMjM}.$$

يوضح الجدول ١٥-٢ خطوات خوارزمية الانحدار الرتيبة، على افتراض أنه لا يوجد تعادل (قيم متساوية) بين قيم $\delta_{ij}s$. في الخطوة ٢ من خوارزمية الانحدار الرتيبة، يتم حساب $\hat{d}_{Bm}$ للكتلة B_m باستخدام متوسط وقيم $\hat{d}_{ij}s$ في B_m:

$$\hat{d}_{B_m} = \sum_{d_{ij} \in B_m} \frac{d_{ij}}{N_m} \qquad (٨\text{-}١٥)$$

حيث N_m هو عدد قيم $d_{ij}s$ في B_m. إذا كان لـ B_m فقط قيمة واحدة من d_{ij}، فإن d_{ij} $= \hat{d}_{imjm}$.

في الخطوة ١ من خوارزمية الانحدار الرتيبة، إذا كان هناك تعادلاً في القيم بين $\delta_{ij}s$، يتم ترتيب $\delta_{ij}s$ ذات القيمة المتساوية في ترتيب متصاعد حسب قيم نظيراتها $d_{ij}s$ في فضاء q من الأبعاد (*Kruskal, 1964a,b*). هناك طريقة أخرى للتعامل مع تعادل القيم بين $\delta_{ij}s$ وذلك بجعل هذه القيم المتساوية لـ $\delta_{ij}s$ تشكِّل كتلة واحدة مع ما يناظرها من قيم $d_{ij}s$ في هذه الكتلة.

بعد استخدام طريقة الانحدار الرتيبة للحصول على قيم $\hat{d}_{ij}s$، نقوم باستخدام المعادلة ١٥-٣ لحساب جهد التهيئة في الخطوة ٣ من خوارزمية *MDS*. كلما كانت قيمة S أصغر، كان أفضل للتهيئة أن تحافظ على نظام ترتيب الاختلافات في المعادلة ١٥-١. يعد كروسكال (*Kruskal, 1964a,b*) أن قيمة S المساوية لـ (٢٠٪) تدل على ضعف تمثيل ومطابقة التهيئة لترتيب الاختلاف في المعادلة ١٥-١، وقيمة S المساوية (١٠٪) تدل على تمثيل ومطابقة مقبولة، وقيمة S المساوية (٥٪) تدل على جودة التمثيل والمطابقة، وقيمة S المساوية (٢,٥٪) تشير إلى تمثيل ومطابقة ممتازة، وقيمة S المساوية لصفر(٠٪) تدل على أفضل تمثيل ومطابقة.

الجدول (١٥-٢)

خوارزمية الاتحاد الرتيبة - (إنجليزي وعربي)

Step	Description
1	Arrange δ_{imjm}, $m = 1, \ldots, M$, in the order from the smallest to the largest
2	Generate the initial M blocks in the same order in Step 1, $B_1, \ldots, B_M$, such that each block, B_m, has only one dissimilarity value, d_{imjm}, and compute $\hat{d}_B$ using Equation 15.8
3	Make the lowest block the active block, and also make it up-active; denote B as the active block, B_- as the next lower block of B, B_+ as the next higher block of B
4	WHILE the active block B is not the highest block
5	IF $\hat{d}_{B_-} < \hat{d}_B < \hat{d}_{B_+}$ /* B is both down-satisfied and up-satisfied, note that the lowest clock is already down-satisfied and the highest block is already up-satisfied */
6	Make the next higher block of B the active block, and make it up-active
7	ELSE
8	IF B is up-active
9	IF $\hat{d}_B < \hat{d}_{B_+}$ /* B is up-satisfied */
10	Make B down-active
11	ELSE
12	Merge B and B_+ to form a new larger block which replaces B and B_+
13	Make the new block as the active block and it is down-active
14	ELSE /* B is down-active */
15	IF $\hat{d}_{B_-} < \hat{d}_B$ /* B is down-satisfied */
16	Make B up-active
17	ELSE
18	Merge B_- and B to form a new larger block which replaces B_- and B
19	Make the new block as the active block and it is up-active
20	$\hat{d}_{ij} = \hat{d}_B$, for each $d_{ij} \in B$ and for each block B in the final sequence of the blocks

الخطوة	الوصف
١	رتب، δ_{imjm}، $m=1, \ldots, M$، ترتيباً تصاعدياً من الأصغر إلى الأكبر.
٢	قم بتوليد عدد M من الكتل (*blocks*) بنفس الترتيب المعمولة به في الخطوة ١، بحيث يكون لدينا الكتل: $B_1, \ldots, B_M$، بحيث تكون لكل كتلة، B_m، قيمة اختلاف واحدة فقط وهي، d_{imjm}، وقم بحساب $\hat{d}_B$ باستخدام المعادلة ١٥-٨.

٣ اجعل الكتلة الأقل هي الكتلة النشطة، واجعلها أيضاً الكتلة فوق النشطة (*up-active*)، نرمز بالرمز B للكتلة النشطة، وبالرمز B_- للكتلة التالية والأقل من B، وبالرمز B_+ للكتلة التالية والأعلى من B.

٤ كرّر (*WHILE*) ما دام أن الكتلة النشطة B ليست هي الكتلة الأعلى.

٥ إذا كان $\hat{d}_{B-} < \hat{d}_B < \hat{d}_{B+}$.

(تعليق: B تكون متحققة من الأسفل (*down- satisfied*) و من الأعلى (*Up-satisfied*)، لاحظ أن الكتلة الأقل هي بالفعل متحققة من الأسفل والكتلة الأعلى أيضاً متحققة من الأعلى).

٦ اجعل الكتلة التالية الأعلى لـ B هي الكتلة النشطة، واجعلها أيضاً فوق النشطة.

٧ خلاف ذلك (*ELSE*).

٨ إذا *(IF)* كانت B هي الكتلة فوق النشطة.

٩ إذا *(IF)* كان $\hat{d}_B < \hat{d}_{B+}$ (مما يعني أن B متحققة من الأعلى).

١٠ أجعل B هي الكتلة تحت النشطة.

١١ خلاف ذلك (*ELSE*).

١٢ ادمج B و B_+ لتشكيل كتلة جديدة أكبر حجماً تستبدل B و B_+.

١٣ اجعل الكتلة الجديدة هي الكتلة النشطة وتكون أيضاً تحت النشطة.

١٤ خلاف ذلك (*ELSE*) (مما يعني أن تكون B تحت النشطة).

١٥ إذا *(IF)* كان $\hat{d}_{B-} < \hat{d}_B$ (مما يعني أن B متحققة من الأسفل).

١٦ اجعل B هي الكتلة فوق النشطة.

١٧ خلاف ذلك (*ELSE*).

١٨ ادمج B_- و B_+ لتشكيل كتلة جديدة أكبر حجماً تستبدل B_- و B.

١٩ اجعل الكتلة الجديدة هي الكتلة النشطة وتكون أيضاً فوق النشطة.

٢٠ $\hat{d}_{ij} = \hat{d}_B$ ، لكل $\hat{d}_{ij} \in B$ ولكل كتلة B في السلسلة الأخيرة من الكتلات.

تقوم الخطوة ٤ من خوارزمية (*MDS*) بتقييم جودة المطابقة (*goodness-of-fit*) باستخدام القيمة S للتهيئة. إذا كانت قيمة S للتهيئة غير مقبولة، تقوم الخطوة ٥ من الخوارزمية بتغيير قيمة التهيئة لتحسين جودة المطابقة باستخدام طريقة الهبوط المتدرج.

تقوم الخطوة ٦ من الخوارزمية بتطبيع متجه كل سجل بيانات في قيمة التهيئة المحدَّثة. تحسب الخطوة ٧ من الخوارزمية القيمة S للتهيئة المحدَّثة.

في الخطوة ٤ من الخوارزمية، يمكن تعيين حد معين *(threshold)* لجودة المطابقة، واستخدامه بحيث تكون قيمة التهيئة مقبولة إذا كانت S للتهيئة أقل من أو يساوي حد جودة المطابقة. ومن ثم، فإن شرط التوقف في الخطوة ٤ من الخوارزمية يظهر بحيث تكون قيمة S أقل من أو تساوي حد جودة المطابقة. إذا كان التغيير في قيمة S صغيراً، بمعنى أنه عندما تبدأ قيمة S في الميل للاستقرار بعد عدة تكرارات من تحديث قيمة التهيئة، ومن ثم فإن إجراء تحديث قيمة التهيئة يمكن إيقافه أيضاً. لذلك فإن تغيير قيمة S، التي هي أصغر من قيمة حد معين، يُعتبر شرط توقف آخر للتكرار يمكن استخدامه في الخطوة ٤ من خوارزمية الـ *MDS*.

إنَّ طريقة الهبوط المتدرج لتحديث التهيئة لـ S في الخطوة ٥ من خوارزمية *MDS* هي طريقة مشابهة لطريقة الهبوط المتدرج المستخدمة لتحديث أوزان الارتباط في طريقة التعلم بالتوالد الخلفي للشبكات العصبية الصناعية (*ANN*) في الفصل ٥. إنَّ الهدف من تحديث قيمة التهيئة، $(x_{11}, \ldots, x_{1q}, \ldots, x_{n1}, \ldots, x_{nq})$ هو تقليل جهد التهيئة في المعادلة ١٥-٣ والتي تظهر فيما يلي:

$$S = \sqrt{\frac{\sum_{ij}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{ij} d_{ij}^2}} = \sqrt{\frac{S^*}{T^*}}, \qquad (٩\text{-}١٥)$$

حيث:

$$S^* = \sum_{ij}\left(d_{ij} - \hat{d}_{ij}\right)^2 \qquad (١٠\text{-}١٥)$$

$$T^* = \sum_{ij} d_{ij}^2. \qquad (١١\text{-}١٥)$$

باستخدام طريقة الهبوط المتدرج، نقوم بتحديث كل x_{kl} حيث أنَّ: $k=1, \ldots, n,$ و $l=1, \ldots, q$، في التهيئة على النحو التالي (*Kruskal, 1964a,b*):

$$x_{kl}(t+1) = x_{kl}(t) + \alpha\Delta x_{kl} = x_{kl}(t) + \alpha(g_{kl}) \Big/ \left(\frac{\sqrt{\sum_{k,l} g_{kl}^2}}{\sum_{k,l} x_{kl}^2} \right), \quad (١٢-١٥)$$

حيث إن:

$$g_{kl} = -\frac{\partial S}{\partial x_{kl}}, \quad (١٣-١٥)$$

و α هي معدل التعلم. وللحصول على قيمة مطبعة لـ x تصبح المعادلة ١٥-١٢:

$$x_{kl}(t+1) = x_{kl}(t) + \alpha\Delta x_{kl} = x_{kl}(t) + \alpha \frac{g_{kl}}{\sqrt{\frac{\sum_{k,l} g_{kl}^2}{n}}} \quad (١٤-١٥)$$

يقدم كروسكال (*Kruskal, 1964a,b*) الصيغة التالية لحساب g_{kl} إذا تمّ حساب قيمة d_{ij} باستخدام المسافة المترية r لمينكوسكي (*Minkowski r-metric distance*):

$$g_{kl} = -\frac{\partial S}{\partial x_{kl}} = S \sum_{i,j} \left[(\rho^{ki} - \rho^{kj}) \left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right) \left(\frac{|x_{il} - x_{jl}|^{r-1}}{d_{ij}^{r-1}} \right) \text{sign}(x_{il} - x_{jl}) \right],$$

(١٥-١٥)

حيث إن:

$$\rho^{ki} = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{if } k \neq i \end{cases} \quad (١٦-١٥)$$

$$\text{sign}(x_{il} - x_{jl}) = \begin{cases} 1 & \text{if } x_{il} - x_{jl} > 0 \\ -1 & \text{if } x_{il} - x_{jl} < 0 \\ 0 & \text{if } x_{il} - x_{jl} = 0 \end{cases} \qquad (١٧\text{-}١٥)$$

إذا كانت $r=2$ في الصيغة ١٥-١٣، وهذا يعني أنه يتم استخدام المسافة الإقليدية لحساب dij.

$$g_{kl} = S\sum_{i,j}\left[(\rho^{ki} - \rho^{kj})\left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*}\right)\left(\frac{x_{il} - x_{jl}}{d_{ij}}\right)\right]. \qquad (١٨\text{-}١٥)$$

مثال ١٥-١:

يوضح الجدول ١٥-٣ ثلاثة سجلات بيانات لتسعة متغيرات جودة، والتي هي جزء من الجدول ٨-١. كما يوضح الجدول ١٥-٤ المسافة الإقليدية لكل زوج من سجلات البيانات الثلاثة في فضاء تُساعي الأبعاد. يتم أخذ هذه المسافة الإقليدية الخاصة بزوج سجلات بيانات، x_i و x_j باعتبارها δ_{ij}. قم بتنفيذ خوارزمية القياس المتعدد الأبعاد (*MDS*) لمجموعة البيانات هذه مع تكرار واحد فقط لتحديث التهيئة لـ $q = 2$، وشرط التوقف $0.2 \geq S$، $\alpha = 5\%$.

في مجموعة البيانات هذه، يوجد ثلاث سجلات بيانات،$n=3$، في فضاء تساعي الأبعاد. لدينا $\delta_{12}=2.65,\ \delta_{13}=2.65,\ \delta_{23}=2$. في الخطوة ١ من خوارزمية *MDS* الموضحة في الجدول ١٥-١، نقوم بتوليد تهيئة أولية لسجلات البيانات الثلاثة في الفضاء ثنائي الأبعاد:

$$x_1 = (1,1) \qquad x_2 = (0,1) \qquad x_3 = (1,0.5).$$

في الخطوة ٢ من خوارزمية *MDS*، نقوم بتطبيع كل سجل بيانات بحيث يحتوي على وحدة الطول، وذلك باستخدام الصيغة ١٥-٢:

$$x_1 = \left(\frac{x_{11}}{\sqrt{x_{11}^2 + x_{12}^2}}, \frac{x_{12}}{\sqrt{x_{11}^2 + x_{12}^2}}\right) = \left(\frac{1}{\sqrt{1^2 + 1^2}}, \frac{1}{\sqrt{1^2 + 1^2}}\right)$$
$$= (0.71, 0.71)$$

$$x_2 = \left(\frac{x_{21}}{\sqrt{x_{21}^2 + x_{22}^2}}, \frac{x_{22}}{\sqrt{x_{21}^2 + x_{22}^2}}\right) = \left(\frac{0}{\sqrt{0^2 + 1^2}}, \frac{1}{\sqrt{0^2 + 1^2}}\right) = (0, 1)$$

$$x_3 = \left(\frac{x_{31}}{\sqrt{x_{31}^2 + x_{32}^2}}, \frac{x_{32}}{\sqrt{x_{31}^2 + x_{32}^2}}\right) = \left(\frac{1}{\sqrt{1^2 + 0.5^2}}, \frac{0.5}{\sqrt{1^2 + 0.5^2}}\right)$$
$$= (0.89, 0.45).$$

الجدول (١٥-٣)

مجموعة البيانات لنظام اكتشاف الأعطال مع ثلاث حالات من الأعطال الآلية الأحادية

متغيرات الخاصية عن جودة وحدات المنتج Attribute Variables about Quality of Parts									رقم الحالة - Instance (الآلة المعطلة - Faulty Machine)
x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	
1	0	1	0	1	0	0	0	1	1 (M1)
0	1	0	0	0	1	0	1	0	2(M2)
0	1	1	1	0	1	1	0	0	3(M3)

الجدول (١٥-٤)

المسافة الإقليدية لكل زوج من سجلات البيانات

$C_3 = \{x_3\}$	$C_2 = \{x_2\}$	$C_1 = \{x_1\}$	
2.65	2.65		$C_1=\{x_1\}$
2			$C_2=\{x_2\}$
			$C_3=\{x_3\}$

يتم حساب المسافة بين كل زوج من سجلات البيانات الثلاثة في الفضاء ثنائي الأبعاد باستخدام إحداثياتها الأولية:

$$d_{12} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2}$$
$$= \sqrt{(0.71 - 0)^2 + (0.71 - x_{22})^2} = 0.77$$

$$d_{13} = \sqrt{(x_{11} - x_{31})^2 + (x_{12} - x_{32})^2}$$
$$= \sqrt{(0.71 - 0.89)^2 + (0.71 - 0.45)^2} = 0.32$$

$$d_{23} = \sqrt{(x_{21} - x_{31})^2 + (x_{22} - x_{32})^2}$$
$$= \sqrt{(0 - 0.89)^2 + (1 - 0.45)^2} = 1.05.$$

قبل أن نقوم بحساب جهد التهيئة الأولية باستخدام الصيغة ١٥-٣، نحتاج إلى استخدام خوارزمية الانحدار الرتيبة في الجدول ١٥-٢ لحساب $\hat{d}_{ij}$. في الخطوة ١ من خوارزمية الانحدار الرتيبة، نقوم بترتيب δ_{imjm}، حيث $m=1, \ldots, M$، ترتيباً تصاعدياً، من الأصغر إلى الأكبر، حيث $M=3$:

$$\delta_{23} < \delta_{12} = \delta_{13}.$$

ولأنه يوجد تعادل بين δ_{12} وδ_{13}، فإن δ_{12} وδ_{13} يتم ترتيبها تصاعدياً بناءً على قيم $d_{12}= 0.77$ و $d_{13}=0.32$:

$$\delta_{23} < \delta_{13} < \delta_{12}.$$

في الخطوة ٢ من خوارزمية الانحدار الرتيبة، نقوم بتوليد الكتل (*Blocks*) الأولية بعدد M بنفس الترتيب في الخطوة ١، B_m، ...، B_1، بحيث يكون لكل كتلة، B_m قيمة اختلاف واحدة فقط، d_{imim}:

$$B_1 = \{d_{23}\} \quad B_2 = \{d_{13}\} \quad B_3 = \{d_{12}\}.$$

نقوم بحساب $\hat{d}_B$ باستخدام الصيغة ١٥-٨:

$$\hat{d}_{B_1} = \sum_{d_{ij} \in B_1} \frac{d_{ij}}{n_1} = \frac{d_{23}}{1} = 1.05$$

$$\hat{d}_{B_2} = \sum_{d_{ij} \in B_2} \frac{d_{ij}}{n_2} = \frac{d_{13}}{1} = 0.32$$

$$\hat{d}_{B_3} = \sum_{d_{ij} \in B_3} \frac{d_{ij}}{n_3} = \frac{d_{12}}{1} = 0.77$$

في الخطوة ٣ من خوارزمية الانحدار الرتيبة، نجعل الكتلة الأقل، B_1، هي الكتلة النشطة:

$$B = B_1 \quad B_- = \emptyset \quad B_+ = B_2,$$

ونجعل B هي الكتلة فوق النشطة. وفي الخطة ٤ من خوارزمية الانحدار الرتيبة، نقوم بالتحقق من أن الكتلة النشطة B_1 ليست هي الكتلة الأعلى. في الخطوة ٥ من خوارزمية الانحدار الرتيبة، نقوم بالتحقق من أن $\hat{d}_B > \hat{d}_{B+}$، ومن ثم لا تكون B مستوفاةً من الأعلى. نذهب إلى الخطوة ٨ من خوارزمية الانحدار الرتيبة ونقوم بالتحقق من أن B نشطة من الأعلي. في الخطوة ٩ من خوارزمية الانحدار الرتيبة، نقوم بالتحقق من أن $\hat{d}_B > \hat{d}_{B+}$، ومن ثم لا تكون B مستوفاةً من أعلى. نذهب إلى الخطوة ١٢ ونقوم بدمج B و$B+$ لتشكيل كتلة أكبر جديدة لتحل محل B_1 وB_2:

$$B_{12} = \{d_{23}, d_{13}\}$$

$$\hat{d}_{B_{12}} = \sum_{d_{ij}\in B_{12}} \frac{d_{ij}}{n_{12}} = \frac{d_{23}+d_{13}}{2} = \frac{1.05+0.32}{2} = 0.69$$

$$B_{12} = \{d_{23}, d_{13}\} \quad B_3 = \{d_{12}\}$$

$$\hat{d}_{B_3} = \sum_{d_{ij}\in B_3} \frac{d_{ij}}{n_3} = \frac{d_{12}}{1} = 0.77.$$

في الخطوة ١٣ من خوارزمية الانحدار الرتيبة، نجعل الكتلة الجديدة B_{12} هي الكتلة النشطة ونجعلها كذلك الكتلة تحت النشطة:

$$B = B_{12} \quad B_- = \emptyset \quad B_+ = B_3.$$

بالعودة إلى الخطوة ٤، نقوم بالتحقق من أن الكتلة النشطة B_{12} ليست هي الكتلة الأعلى. في الخطوة ٥، نقوم بالتحقق من أن B مستوفاة أو متحققة من الأعلى مع $\hat{d}_{12} < \hat{d}_3$ وأيضاً مستوفاة من الأسفل. لذا، نقوم بتنفيذ الخطوة ٦ لجعل B_3 هي الكتلة النشطة ولجعلها فوق النشطة:

$$B_{12} = \{d_{23}, d_{13}\} \quad B_3 = \{ d_{12}\}$$

$$\hat{d}_{B_{12}} = \sum_{d_{ij}\in B_{12}} \frac{d_{ij}}{n_{12}} = \frac{d_{23}+d_{13}}{2} = \frac{1.05+0.32}{2} = 0.69$$

$$\hat{d}_{B_3} = \sum_{d_{ij}\in B_3} \frac{d_{ij}}{n_3} = \frac{d_{12}}{1} = 0.77.$$

$$B = B_3 \quad B_- = B_{12} \quad B_+ = \emptyset.$$

بالعودة إلى الخطوة ٤ مرة أخرى، نقوم بالتحقق من أن تلك الكتلة النشطة B هي الكتلة الأعلى، نقوم بالخروج من تعليمة التكرار *(WHILE)*، وبتنفيذ الخطوة ٢٠ وهي الخطوة الأخيرة من خوارزمية الانحدار الرتيبة، وإسناد القيم التالية الخاصة بـ $\hat{d}_{ij}s$:

$$\hat{d}_{12} = \hat{d}_{B_3} = 0.77$$

$$\hat{d}_{13} = \hat{d}_{B_{12}} = 0.69$$

$$\hat{d}_{23} = \hat{d}_{B_{12}} = 0.69.$$

وبقيم $\hat{d}_{ij}$ وقيم d_{ij}:

$$d_{12} = 0.77$$

$$d_{13} = 0.32$$

$$d_{23} = 1.05,$$

نقوم الآن بتنفيذ الخطوة ٣ من خوارزمية *MDS* لحساب جهد التهيئة الأوّلي باستخدام المعادلات ١٥-٩ وحتى ١٥-١١:

$$S^* = \sum_{ij} \left(d_{ij} - \hat{d}_{ij}\right)^2$$
$$= (0.77 - 0.77)^2 + (0.32 - 0.69)^2 + (1.05 - 0.69)^2 = 0.27$$

$$T^* = \sum_{ij} d_{ij}^2 = 0.77^2 + 0.32^2 + 1.05^2 = 0.61$$

$$S = \sqrt{\frac{S^*}{T^*}} = \sqrt{\frac{0.27}{0.61}} = 0.67.$$

هذا المستوى من الجهد يشير إلى ضعف جودة المطابقة (*goodness-of-fit*). في الخطوة ٤ من خوارزمية *MDS*، نقوم بالتحقق من أن S لا تحقق شرط توقف تعليمة التكرار (*REPEAT*). في الخطوة ٥ من خوارزمية *MDS*، نقوم بتحديث التهيئة باستخدام المعادلات ١٥-١٤، ١٥-١٦ و١٥-١٨ مع $k = 1,2,3$ و$l = 1,2$:

$$\begin{aligned}
g_{kl} = g_{11} &= S \sum_{i,j} \left[(\rho^{ki} - \rho^{kj}) \left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right) \left(\frac{x_{il} - x_{jl}}{d_{ij}} \right) \right] \\
&= (0.67) \sum_{i,j} \left[(\rho^{1i} - \rho^{1j}) \left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right) \left(\frac{x_{i1} - x_{j1}}{d_{ij}} \right) \right] \\
&= (0.67) \left[(\rho^{11} - \rho^{12}) \left(\frac{d_{12} - \hat{d}_{12}}{S^*} - \frac{d_{12}}{T^*} \right) \left(\frac{x_{11} - x_{21}}{d_{12}} \right) \right. \\
&\quad + (\rho^{11} - \rho^{13}) \left(\frac{d_{13} - \hat{d}_{13}}{S^*} - \frac{d_{13}}{T^*} \right) \left(\frac{x_{11} - x_{31}}{d_{13}} \right) \\
&\quad \left. + (\rho^{12} - \rho^{13}) \left(\frac{d_{23} - \hat{d}_{23}}{S^*} - \frac{d_{23}}{T^*} \right) \left(\frac{x_{21} - x_{31}}{d_{23}} \right) \right] \\
&= (0.67) \left[(1 - 0) \left(\frac{0.77 - 0.77}{0.27} - \frac{0.77}{0.61} \right) \left(\frac{0.71 - 0}{0.77} \right) \right. \\
&\quad + (1 - 0) \left(\frac{0.32 - 0.69}{0.27} - \frac{0.32}{0.61} \right) \left(\frac{0.71 - 0.89}{0.32} \right) \\
&\quad \left. + (0 - 0) \left(\frac{1.05 - 0.69}{0.27} - \frac{1.05}{0.61} \right) \left(\frac{0 - 0.89}{1.05} \right) \right] \\
&= -0.13
\end{aligned}$$

$$g_{kl} = g_{12} = S\sum_{i,j}\left[(\rho^{ki} - \rho^{kj})\left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*}\right)\left(\frac{x_{il} - x_{jl}}{d_{ij}}\right)\right]$$

$$= (0.67)\sum_{i,j}\left[(\rho^{1i} - \rho^{1j})\left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*}\right)\left(\frac{x_{i2} - x_{j2}}{d_{ij}}\right)\right]$$

$$= (0.67)\left[(\rho^{11} - \rho^{12})\left(\frac{d_{12} - \hat{d}_{12}}{S^*} - \frac{d_{12}}{T^*}\right)\left(\frac{x_{12} - x_{22}}{d_{12}}\right)\right.$$

$$+ (\rho^{11} - \rho^{13})\left(\frac{d_{13} - \hat{d}_{13}}{S^*} - \frac{d_{13}}{T^*}\right)\left(\frac{x_{12} - x_{32}}{d_{13}}\right)$$

$$\left. + (\rho^{12} - \rho^{13})\left(\frac{d_{23} - \hat{d}_{23}}{S^*} - \frac{d_{23}}{T^*}\right)\left(\frac{x_{22} - x_{32}}{d_{23}}\right)\right]$$

$$= (0.67)\left[(1 - 0)\left(\frac{0.77 - 0.77}{0.27} - \frac{0.77}{0.61}\right)\left(\frac{0.71 - 1}{0.77}\right)\right.$$

$$+ (1 - 0)\left(\frac{0.32 - 0.69}{0.27} - \frac{0.32}{0.61}\right)\left(\frac{0.71 - 0.45}{0.32}\right)$$

$$\left. + (0 - 0)\left(\frac{1.05 - 0.69}{0.27} - \frac{1.05}{0.61}\right)\left(\frac{1 - 0.45}{1.05}\right)\right]$$

$$= -0.71$$

$$g_{kl} = g_{21} = S\sum_{i,j}\left[(\rho^{ki} - \rho^{kj})\left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*}\right)\left(\frac{x_{il} - x_{jl}}{d_{ij}}\right)\right]$$

$$= (0.67)\sum_{i,j}\left[(\rho^{2i} - \rho^{2j})\left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*}\right)\left(\frac{x_{i1} - x_{j1}}{d_{ij}}\right)\right]$$

$$= (0.67)\left[(\rho^{21} - \rho^{22})\left(\frac{d_{12} - \hat{d}_{12}}{S^*} - \frac{d_{12}}{T^*}\right)\left(\frac{x_{11} - x_{21}}{d_{12}}\right)\right.$$

$$+ (\rho^{21} - \rho^{23}) \left(\frac{d_{13} - \hat{d}_{13}}{S^*} - \frac{d_{13}}{T^*} \right) \left(\frac{x_{11} - x_{31}}{d_{13}} \right)$$

$$+ (\rho^{22} - \rho^{23}) \left(\frac{d_{23} - \hat{d}_{23}}{S^*} - \frac{d_{23}}{T^*} \right) \left(\frac{x_{21} - x_{31}}{d_{23}} \right) \Bigg]$$

$$= (0.67) \Bigg[(0 - 1) \left(\frac{0.77 - 0.77}{0.27} - \frac{0.77}{0.61} \right) \left(\frac{0.71 - 0}{0.77} \right)$$

$$+ (0 - 0) \left(\frac{0.32 - 0.69}{0.27} - \frac{0.32}{0.61} \right) \left(\frac{0.71 - 0.89}{0.32} \right)$$

$$+ (1 - 0) \left(\frac{1.05 - 0.69}{0.27} - \frac{1.05}{0.61} \right) \left(\frac{0 - 0.89}{1.05} \right) \Bigg]$$

$$= 1.07$$

$$g_{kl} = g_{22} = S \sum_{i,j} \left[(\rho^{ki} - \rho^{kj}) \left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right) \left(\frac{x_{il} - x_{jl}}{d_{ij}} \right) \right]$$

$$= (0.67) \sum_{i,j} \left[(\rho^{2i} - \rho^{2j}) \left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right) \left(\frac{x_{i2} - x_{j2}}{d_{ij}} \right) \right]$$

$$= (0.67) \Bigg[(\rho^{21} - \rho^{22}) \left(\frac{d_{12} - \hat{d}_{12}}{S^*} - \frac{d_{12}}{T^*} \right) \left(\frac{x_{12} - x_{22}}{d_{12}} \right)$$

$$+ (\rho^{21} - \rho^{23}) \left(\frac{d_{13} - \hat{d}_{13}}{S^*} - \frac{d_{13}}{T^*} \right) \left(\frac{x_{11} - x_{31}}{d_{13}} \right)$$

$$+ (\rho^{22} - \rho^{23}) \left(\frac{d_{23} - \hat{d}_{23}}{S^*} - \frac{d_{23}}{T^*} \right) \left(\frac{x_{22} - x_{32}}{d_{23}} \right) \Bigg]$$

$$= (0.67) \Bigg[(0 - 1) \left(\frac{0.77 - 0.77}{0.27} - \frac{0.77}{0.61} \right) \left(\frac{0.71 - 1}{0.77} \right)$$

$$+ (0-0)\left(\frac{0.32-0.69}{0.27}-\frac{0.32}{0.61}\right)\left(\frac{0.71-0.45}{0.32}\right)$$

$$+ (1-0)\left(\frac{1.05-0.69}{0.27}-\frac{1.05}{0.61}\right)\left(\frac{1-0.45}{1.05}\right)\Bigg]$$

$$= -0.45$$

$$g_{kl} = g_{31} = S\sum_{i,j}\left[(\rho^{ki}-\rho^{kj})\left(\frac{d_{ij}-\hat{d}_{ij}}{S^*}-\frac{d_{ij}}{T^*}\right)\left(\frac{x_{il}-x_{jl}}{d_{ij}}\right)\right]$$

$$= (0.67)\sum_{i,j}\left[(\rho^{3i}-\rho^{3j})\left(\frac{d_{ij}-\hat{d}_{ij}}{S^*}-\frac{d_{ij}}{T^*}\right)\left(\frac{x_{i1}-x_{j1}}{d_{ij}}\right)\right]$$

$$= (0.67)\Bigg[(\rho^{31}-\rho^{32})\left(\frac{d_{12}-\hat{d}_{12}}{S^*}-\frac{d_{12}}{T^*}\right)\left(\frac{x_{11}-x_{21}}{d_{12}}\right)$$

$$+ (\rho^{31}-\rho^{33})\left(\frac{d_{13}-\hat{d}_{13}}{S^*}-\frac{d_{13}}{T^*}\right)\left(\frac{x_{11}-x_{31}}{d_{13}}\right)$$

$$+ (\rho^{32}-\rho^{33})\left(\frac{d_{23}-\hat{d}_{23}}{S^*}-\frac{d_{23}}{T^*}\right)\left(\frac{x_{21}-x_{31}}{d_{23}}\right)\Bigg]$$

$$= (0.67)\Bigg[(0-0)\left(\frac{0.77-0.77}{0.27}-\frac{0.77}{0.61}\right)\left(\frac{0.71-0}{0.77}\right)$$

$$+ (0-1)\left(\frac{0.32-0.69}{0.27}-\frac{0.32}{0.61}\right)\left(\frac{0.71-0.89}{0.32}\right)$$

$$+ (0-1)\left(\frac{1.05-0.69}{0.27}-\frac{1.05}{0.61}\right)\left(\frac{0-0.89}{1.05}\right)\Bigg]$$

$$= 0.90$$

$$g_{kl} = g_{32} = S\sum_{i,j}\left[(\rho^{kl}-\rho^{kj})\left(\frac{d_{ij}-\hat{d}_{ij}}{S^*}-\frac{d_{ij}}{T^*}\right)\left(\frac{x_{il}-x_{jl}}{d_{ij}}\right)\right]$$

$$= (0.67)\sum_{i,j}\left[(\rho^{3i} - \rho^{3j})\left(\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*}\right)\left(\frac{x_{i2} - x_{j2}}{d_{ij}}\right)\right]$$

$$= (0.67)\left[(\rho^{31} - \rho^{32})\left(\frac{d_{12} - \hat{d}_{12}}{S^*} - \frac{d_{12}}{T^*}\right)\left(\frac{x_{12} - x_{22}}{d_{12}}\right)\right.$$

$$+ (\rho^{31} - \rho^{33})\left(\frac{d_{13} - \hat{d}_{13}}{S^*} - \frac{d_{13}}{T^*}\right)\left(\frac{x_{12} - x_{32}}{d_{13}}\right)$$

$$\left. + (\rho^{32} - \rho^{33})\left(\frac{d_{23} - \hat{d}_{23}}{S^*} - \frac{d_{23}}{T^*}\right)\left(\frac{x_{22} - x_{32}}{d_{23}}\right)\right]$$

$$= (0.67)\left[(0 - 0)\left(\frac{0.77 - 0.77}{0.27} - \frac{0.77}{0.61}\right)\left(\frac{0.71 - 1}{0.77}\right)\right.$$

$$+ (0 - 1)\left(\frac{0.32 - 0.69}{0.27} - \frac{0.32}{0.61}\right)\left(\frac{0.71 - 0.45}{0.32}\right)$$

$$\left. + (0 - 1)\left(\frac{1.05 - 0.69}{0.27} - \frac{1.05}{0.61}\right)\left(\frac{1 - 0.45}{1.05}\right)\right]$$

$$= 0.77$$

$$x_{kl}(t + 1) = x_{kl}(t) + \alpha\Delta x_{kl} = x_{kl}(t) + \alpha\frac{g_{kl}}{\sqrt{\frac{\sum_{k,l} g_{kl}^2}{n}}}$$

$$x_{11}(1) = x_{11}(0) + 0.2\frac{g_{11}}{\sqrt{\frac{g_{11}^2 + g_{12}^2 + g_{21}^2 + g_{22}^2 + g_{31}^2 + g_{32}^2}{3}}}$$

$$= 0.71 + 0.2\frac{-0.13}{\sqrt{\frac{(-0.13)^2 + (-0.17)^2 + 1.07^2 + (-0.45)^2 + 0.90^2 + 0.77^2}{3}}} = 0.70$$

$$x_{12}(1) = x_{12}(0) + 0.2 \frac{g_{12}}{\sqrt{\frac{g_{11}^2 + g_{12}^2 + g_{21}^2 + g_{22}^2 + g_{31}^2 + g_{32}^2}{3}}}$$

$$= 0.71 + 0.2 \frac{-0.71}{\sqrt{\frac{(-0.13)^2 + (-0.17)^2 + 1.07^2 + (-0.45)^2 + 0.90^2 + 0.77^2}{3}}} = 0.63$$

$$x_{21}(1) = x_{21}(0) + 0.2 \frac{g_{21}}{\sqrt{\frac{g_{11}^2 + g_{12}^2 + g_{21}^2 + g_{22}^2 + g_{31}^2 + g_{32}^2}{3}}}$$

$$= 0 + 0.2 \frac{1.07}{\sqrt{\frac{(-0.13)^2 + (-0.17)^2 + 1.07^2 + (-0.45)^2 + 0.90^2 + 0.77^2}{3}}} = 0.12$$

$$x_{22}(1) = x_{22}(0) + 0.2 \frac{g_{22}}{\sqrt{\frac{g_{11}^2 + g_{12}^2 + g_{21}^2 + g_{22}^2 + g_{31}^2 + g_{32}^2}{3}}}$$

$$= 1 + 0.2 \frac{-0.45}{\sqrt{\frac{(-0.13)^2 + (-0.17)^2 + 1.07^2 + (-0.45)^2 + 0.90^2 + 0.77^2}{3}}} = 0.95$$

$$x_{31}(1) = x_{31}(0) + 0.2 \frac{g_{31}}{\sqrt{\frac{g_{11}^2 + g_{12}^2 + g_{21}^2 + g_{22}^2 + g_{31}^2 + g_{32}^2}{3}}}$$

$$= 0.89 + 0.2 \frac{0.90}{\sqrt{\frac{(-0.13)^2 + (-0.17)^2 + 1.07^2 + (-0.45)^2 + 0.90^2 + 0.77^2}{3}}} = 0.99$$

$$x_{32}(1) = x_{32}(0) + 0.2 \frac{g_{32}}{\sqrt{\frac{g_{11}^2 + g_{12}^2 + g_{21}^2 + g_{22}^2 + g_{31}^2 + g_{32}^2}{3}}}$$

$$= 0.45 + 0.2 \frac{0.77}{\sqrt{\frac{(-0.13)^2 + (-0.17)^2 + 1.07^2 + (-0.45)^2 + 0.90^2 + 0.77^2}{3}}} = 0.54.$$

ومن ثم، بعد تحديث التهيئة الأولية في الخطوة ٥ من خوارزمية *MDS*، فإننا نحصل على:

$$x_1 = (0.70, 0.63) \qquad x_2 = (0.12, 0.95) \qquad x_3 = (0.99, 0.54).$$

في الخطوة ٦ من خوارزمية *MDS*، نقوم بتطبيع كل x_i:

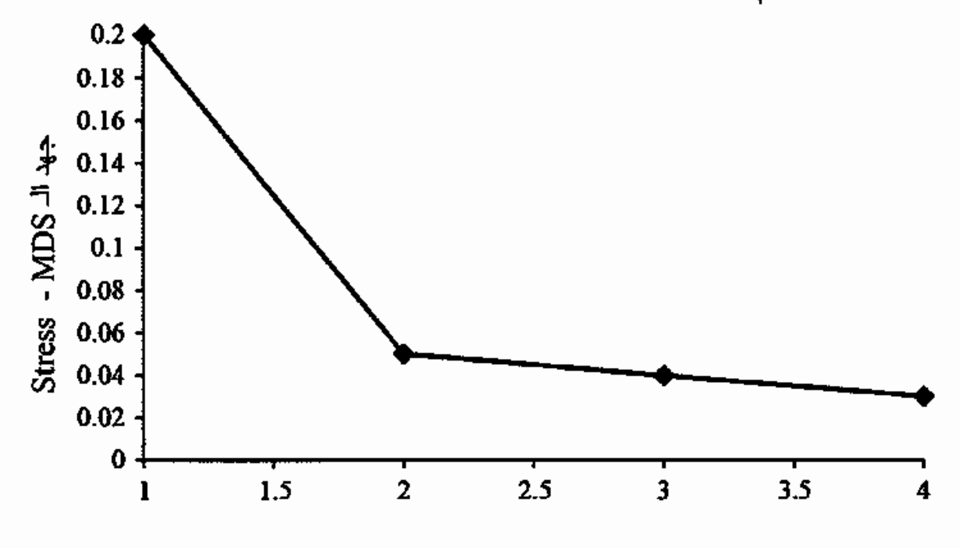
$$x_1 = \left(\frac{0.70}{\sqrt{0.70^2 + 0.63^2}}, \frac{0.63}{\sqrt{0.70^2 + 0.63^2}}\right) = (0.74, 0.67)$$

$$x_2 = \left(\frac{0.12}{\sqrt{0.12^2 + 0.95^2}}, \frac{0.95}{\sqrt{0.12^2 + 0.95^2}}\right) = (0.13, 0.99)$$

$$x_3 = \left(\frac{0.99}{\sqrt{0.99^2 + 0.54^2}}, \frac{0.54}{\sqrt{0.99^2 + 0.54^2}}\right) = (0.88, 0.48).$$

الشكل (١٥-١)

مثال على رسم الجهد الخاص بنتيجة القياس المتعدد الأبعاد *(MDS)* مقابل عدد الأبعاد



١٥-٢ عدد الأبعاد (Number of Dimensions):

تبدأ خوارزمية القياس المتعدد الأبعاد (*MDS*) في الجزء ١٥-١ بالقيمة المعطاة *q* وهي تمثل عدد الأبعاد. قبل الحصول على النتيجة النهائية *MDS* لمجموعة بيانات، ينصح باستخدام عدة قيم لـ *q* للحصول على نتيجة الـ *MDS* لكل قيمة *q*، ومن ثمّ نقوم بعمل رسم بياني لجهد التهيئة مقابل قيمة *q*، ونقوم باختيار قيمة *q* من الرسم البياني عند النقطة التي يحدث فيها انعطاف واضح على شكل كوع الذراع واختيار القيمة المقابلة لنتيجة (*MDS*). الشكل ١٥-١ يوضح رسماً بيانياً للجهد مقابل *q*. وتكون قيمة *q* عند المنعطف في هذا الرسم هي ٢. يتم اختيار قيمة *q* عند المنعطف، وذلك لأن الجهد يتحسن كثيراً قبل نقطة المنعطف ولكنه يستقر بعد نقطة المنعطف. على سبيل المثال. في الدراسة التي أجراها يي (*Ye, 1998*). يتم الحصول على نتائج القياس المتعدد الأبعاد لقيم مختلفة خاصة بـ *q*، *q= 1,2,3,4,5,and 6*. تظهر قيم الجهد لنتائج القياس المتعدد الأبعاد *MDS* أن نقطة المنعطف تكون عند *q=3*.

١٥-٣ قياس الفروقات الفردية للقياس المتعدد الأبعاد الموزون (INDSCALE Weighted MDS):

في الدراسة التي أجراها يي (*Ye, 1998*)، تمّ إعطاء عدد من الأشخاص (وهم يمثلون عينات البحث – *subjects* - مصنّفين كمبرمجين خبراء ومبرمجين مبتدئين) قائمةً تحتوي مفاهيم لغة البرمجة *C* وتمّ الطلب منهم أن يقوموا بتقدير الاختلاف لكل زوج من هذه المفاهيم. ومن ثم، تمّ الحصول على مصفوفة اختلاف لمفاهيم لغة البرمجة *C* من كل عينة بحثية. وباعتبار أن كل مفهوم برمجة يمثل سجل بيانات، تمّ استخدام قياس الفروقات الفردية (*INDSCALE*) في الدراسة لأخذ مصفوفات الاختلاف لسجلات البيانات من العينات البحثية (المبرمجين) كمدخلات ومن ثمّ استخراج المخرجات بما في ذلك التهيئة الخاصة بإحداثيات كل سجل بيانات في فضاء بعدد *q* من الأبعاد للمجموعة الكاملة من المبرمجين ومتجه وزن لكل مبرمج. يحتوي متجه الوزن لمبرمج ما على قيمة وزن لهذا المبرمج في كل بعد.

إنَّ تطبيق متجه الوزن لمبرمج ما على تهيئة مجموعة إحداثيات المفاهيم يعطي تهيئة إحداثيات المفاهيم المأخوذة من المبرمج - يتم تنظيم مفاهيم لغة البرمجة C من قبل كل مبرمج. حيث أن متجهات الوزن المختلفة للمبرمجين الأفراد تعكس اختلافاتهم في تنظيم المعرفة، فإن الدراسة تطبق منهج تباين تحليل الزوايا (*ANAVA*) على متجهات الوزن الخاصة بالمبرمجين الأفراد لتحليل اختلافات الزوايا لمتجهات الوزن وتقييم أهمية اختلافات تنظيم المعرفة بين مجموعتين ممن يملك المهارة، الخبراء والمبتدئون.

وبشكل عام، فإن قياس الفروقات الفردية (*INDSCALE*) أو القياس المتعدد الأبعاد الموزون (*weighted MDS*) يأخذان مصفوفات اختلاف الخاصة بعدد n من الأهداف المبحوثة (*objects*) من عدد m من العينات البحثية وينتجان تهيئة مجموعة إحداثيات الهدف المبحوث:

$$x_i = (x_{i1}, \dots, x_{iq}), \quad i = 1, \dots, n,$$

ومتجهات الوزن للعينات البحثية الفردية:

$$w_j = (w_{j1}, \dots, w_{jq}), \quad j = 1, \dots, m.$$

متجه الوزن لعينة بحثية تعكس البروز النسبي لكل بعُد من فضاء التهيئة للعينة البحثية.

١٥-٤ البرمجيات والتطبيقات (Software and Applications):

يتم دعم القياس المتعدد الأبعاد (*MDS*) بالعديد من حزم البرمجيات الإحصائية، بما في ذلك *SAS MDS* وإجراءات قياس الفروقات الفردية *IDNSCALE (www.sas.com)*. ويرد تطبيق للقياس المتعدد الأبعاد *(MDS)* وقياس الفروقات الفردية (*IDNSCALE*) لتحديد الاختلافات بين الخبراء والمبتدئين في تمثيل المعرفة في الجزء ١٥-٣ بالتفاصيل في يي (*Ye, 1998*).

التمارين (Exercises):

٢-١ استمر في عمل المثال ١٥-١ لتنفيذ التكرار التالي من تحديث التهيئة.

٢-٢ بالنظر إلى مجموعة البيانات المكونة من ثلاثة سجلات لبيانات في الحالات أرقام ٤،٥، و٦ في الجدول ٨-١. استخدم المسافة الإقليدية لكل زوج (x_i و x_j) من سجلات البيانات الثلاثة في فضاء تُساعي الأبعاد بوصفها δ_{ij}. ثم نفّذ القياس المتعدد الأبعاد MDA لمجموعة البيانات هذه مع تكرار واحد فقط لتحديث التهيئة لـ $q=3$، وشرط التوقف $S \leq 5\%, , \alpha = 0.2$.

٢-٣ بالنظر إلى مجموعة البيانات في الجدول ٨-١ المكونة من تسع سجلات بيانات في الحالات ١-٩. استخدم المسافة الإقليدية لكل زوج (x_i و x_j) من سجلات البيانات التسعة في الفضاء تساعي الأبعاد بوصفها δ_{ij}. ثم نفّذ القياس المتعدد الأبعاد MDS لمجموعة البيانات هذه مع تكرار واحد فقط لتحديث التهيئة لـ $q=3$، وشرط التوقف $S \leq 5\%, , \alpha = 0.2$.

الجزء الخامس

خوارزميات استكشاف الأنماط المتطرفة والشاذة

Algorithms for Mining Outlier and Anomaly Patterns

١٦- مخطط التحكم أحادي المتغير
Univariate Control Charts

المتطرف والشاذ هي سجلات بيانات تحيد بشكل كبير عن المعيار الذي تتبعه غالبية سجلات البيانات. قد يعود سبب ظهور السجلات الشاذة والمتطرفة إلى وجود عطل في آلة التصنيع، وبالتالي يتم فَقْد التحكم في عملية التصنيع، أو إلى وجود هجوم عبر الإنترنت بحيث يختلف سلوك الاستخدام إلى حد كبير عن سلوك الاستخدام الطبيعي لأنظمة الحاسوب والشبكات، وهلم جرا. يُعدُّ اكتشاف السجلات والقيم المتطرفة والشاذة أمرًا مهمًا في العديد من المجالات. على سبيل المثال، يعدُّ اكتشاف عملية تصنيع خارجة عن التحكم والسيطرة بسرعة أمرًا مهمًا للحد من تكاليف التصنيع من خلال تجنب إنتاج مزيد من الوحدات التالفة من منتج ما. كما أنَّ الاكتشاف المبكر عن أي هجوم عبر الإنترنت يُعتَبر أمراً حاسمًا لحماية أنظمة الحاسب والشبكة من الخطر.

تعمل تقنيات مخطط التحكم (*Control Chart*) على تعريف واكتشاف المتطرف والشاذ من البيانات على أساس إحصائي. يصف هذا الفصل مخططات التحكم أحادية المتغير التي تراقب متغيرًا واحدًا لغرض اكتشاف الوضع الشاذ. يصف الفصل السابع عشر مخططات التحكم المتعددة المتغيرات التي تراقب متغيرات متعددة في وقت واحد لغرض اكتشاف الوضع الشاذ. تشتمل مخططات التحكم أحادية المتغير الموضحة في هذا الفصل على مخطط التحكم لشوارتز (*Shewhart control charts*)، ومخططات تحكم المجموع التراكمي (*CUSUM*)، ومخططات تحكم المتوسط المتحرك الموزون الأسي (*EWMA*)، ومخططات تحكم الدرجة التراكمية (*cuscore control charts*). وترد قائمة من حزم البرمجيات التي تدعم مخططات التحكم أحادية المتغير. وترد بعض تطبيقات مخططات التحكم أحادية المتغير مع المراجع.

١٦-١ مخططات التحكم لشوارتز (Shewhart Control Charts):

تشتمل مخططات التحكم لشوارتز على مخططات التحكم في المتغير، وكلٌّ منها يراقب متغيراً بالقيم الرقمية (على سبيل المثال، قُطر الثقب الذي تمّ عمله بواسطة آلة قطع معينة)، ومخططات التحكم في خاصية متغير ما، كلٌّ منها يراقب خاصية تلخص قيماً نوعية (على

سبيل المثال، الجزء المعيب وغير المعيب من وحدات الإنتاج). عند رصد عينات من سجلات البيانات، تكون مخططات التحكم بالمتغير، على سبيل المثال، تكون المخططات التالية قابلة للتطبيق: كمخططات التحكم بالمتوسط $\bar{x}$ لاكتشاف الحالات الشاذة المتعلقة بمتوسط (*mean*) عملية ما، ومخططات التحكم بـ R وs لاكتشاف الحالات الشاذة المتعلقة بتباين ما (*variance*). عندما يمكن رصد سجلات بيانات فردية فقط، تكون مخططات التحكم بالمتغير، على سبيل المثال، مخططات التحكم الفردية، قابلة أكثر للتطبيق. بالنسبة إلى مجموعة بيانات بها سجلات بيانات فردية بدلًا من عينات من سجلات البيانات، يكون لكل من مخططي تحكم المجموع التراكمي (*CUSUM*) في الجزء ١٦-٢ ومخططات تحكم المتوسط المتحرك الموزون الأسي (*EWMA*) في الجزء ١٦-٣ مزايا أكثر من مخططات التحكم الفردية.

الجدول (١٦-١)

عينات من ملحوظات البيانات المرصودة

الانحراف المعياري للعينة **Sample Standard Deviation**	متوسط العينة **Sample Mean**	ملحوظات البيانات المرصودة في كل عينة **Data Observations in Each Sample**	العينة **Sample**
S_1	$\bar{x}_1$	$x_{11}, \dots, x_{1j}, \dots, x_{1n}$	1
...	...	...	...
S_i	$\bar{x}_i$	$x_{i1}, \dots, x_{ij}, \dots, x_{in}$	i
...	...	...	...
S_m	$\bar{x}_m$	$x_{m1}, \dots, x_{mj}, \dots, x_{mn}$	m

نقوم بوصف مخططات التحكم بالمتوسط $\bar{x}$ لتوضيح كيفية عمل مخططات التحكم لشوارتز. ليكن لدينا متغير x الذي يأخذ عدد m من العينات لعدد n من ملحوظات البيانات المرصودة والخاصة بعملية ما كما هو مبين في الجدول رقم ١٦-١. يفترض مخطط التحكم بمتوسط العينة $\bar{x}$ أن x موزعة طبيعياً وبمتوسط عينات μ وانحراف معياري للعينات σ عندما تكون العملية تحت التحكم.

يتم حساب قيمة $\bar{x}_i$، وs_i، حيث، $i=1, \dots, m$ ، في الجدول ١٦-١ على النحو التالي:

$$\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n} \qquad (١-١٦)$$

$$S_i = \sqrt{\frac{\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2}{n-1}}. \qquad (٢-١٦)$$

ويتم تقدير قيم المتوسط μ والانحراف المعياري σ باستخدام $\bar{\bar{x}}$ و $\bar{S}$:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{m} \bar{x}_i}{m} \qquad (٣-١٦)$$

$$\bar{S} = \frac{\sum_{i=1}^{m} S_i}{m}. \qquad (٤-١٦)$$

إذا كان حجم العينة n كبيرًا، فإن $\bar{x}_i$ يتْبَع توزيعاً طبيعياً وفقاً لنظرية النهاية المركزية (*central limit theory*). واحتمال أن يقع متوسط العينة $\bar{x}_i$ ضمن ثلاث انحرافات معيارية من متوسط العينات يبلغ حوالي ٩٩،٧٪ استنادًا إلى دالة الكثافة الاحتمالية للتوزيع الطبيعي:

$$P(\bar{\bar{x}} - 3\bar{S} \le \bar{x}_i \le \bar{\bar{x}} + 3\bar{S}) = 99.7\% \qquad (٥-١٦)$$

وحيث إن احتمال أن يقع $\bar{x}_i$ خارج ثلاثة انحرافات معيارية من متوسط العينات هو ٠,٣٪ فقط، فإن متوسط العينة $\bar{x}_i$ هذا يُعتبر متفردًا أو شاذًا وقد يكون ذلك ناجمًا عن عملية خارج السيطرة والتحكم. وبالتالي، عادةً ما يتم استخدام متوسط العينات المقدّر وحدود التحكم المسماة ٣- سيغما (*3-sigma control limits*)، والتي تشير إلى ٣ انحرافات معيارية أعلى أو أقل من متوسط العينات μ باعتبارهما المحور (*centerline*) وحدود التحكم (*Control limits*) (UCL لحد التحكم الأعلى وLCL لحد التحكم الأدنى)، على التوالي، لمتوسط العملية التي تحت السيطرة في مخطط التحكم بمتوسط العينات $\bar{x}$:

$$\text{Centerline} = \bar{\bar{x}} \qquad (٦-١٦)$$

$$UCL = \bar{\bar{x}} + 3\bar{S} \qquad (٧-١٦)$$

$$LCL = \bar{\bar{x}} - 3\bar{S} \qquad (٨-١٦)$$

مخطط التحكم $\bar{x}$ يراقب $\bar{x}_i$ من العينة i الخاصة بملحوظات البيانات المرصودة. إذا وقع $\bar{x}_i$ ضمن النطاق $[UCL, LCL]$، فعليه تُعتبر هذه العملية تحت السيطرة؛ وخلاف ذلك، نعتبر أنه تمّ اكتشاف الشاذ وتُعتبَر العملية خارجة عن السيطرة والتحكم.

باستخدام حدود التحكم ٣- سيغما في مخطط التحكم لـــ $\bar{x}$، لا يزال هناك نسبة احتمال ٠,٣٪ أن تكون العملية تحت السيطرة ولكن تقع ملحوظة البيانات المرصودة خارج حدود السيطرة ويتم توليد إشارة خارج السيطرة *(out-of-control signal)* عن طريق مخطط التحكم لـــ $\bar{x}$. إذا كانت العملية تحت السيطرة ولكن مخطط التحكم يعطي إشارة خارج السيطرة، تكون الإشارة إنذارا خاطئًا. معدل الإنذارات الخاطئة (***rate of false alarm***) هي نسبة عدد الإنذارات الخاطئة إلى العدد الإجمالي لعينات البيانات التي يجري رصدها. إذا كانت العملية خارجة عن السيطرة ومخطط التحكم يولد إشارة خارج السيطرة، يكون لدينا زيارة ناجحة (***hit***). معدل الزيارات الناجحة هو نسبة عدد الزيارات الناجحة إلى العدد الإجمالي من عينات البيانات. باستخدام حدود التحكم ٣- سيغما، ينبغي أن يكون لدينا معدل الزيارة الناجحة ٩٩,٧٪ ومعدل الإنذار الخاطيء ٠,٣٪.

إذا لم يكن حجم العينة n كبيراً، فإن تقدير الانحراف المعياري بواسطة $\bar{S}$ قد يكون بعيدًا إلى حد ما، وربما يحتاج المعامل لـ $\bar{S}$ في المعادلة ١٦- ٧ و١٦- ٨ أن يتم تعديله إلى قيمة مختلفة عن ٣ من أجل وضع حدود تحكم مناسبة حتى تقع الغالبية العظمى من البيانات تحت حدود السيطرة إحصائيا. يعطي مونتغمري (***Montgomery, 2001***) معاملات مناسبة لتحديد حدود التحكم لقيم متنوعة من حجم العينة n.

يُظهِر مخطط التحكم لـــ $\bar{x}$ كيف تعمل مخططات التحكم الإحصائية، مثل مخططات التحكم لشوارتز، علي تأسيس المحور وحدود التحكم على أساس التوزيع الاحتمالي للمتغير المستهدف وتقدير مَعلَمات التوزيع من عينات البيانات. وبشكل عام، يتم تحديد قيمة محور مخطط التحكم مساويةً للقيمة المتوقعة للمتغير، ويتم تحديد حدود التحكم بحيث تقع الغالبية العظمى من البيانات في حدود التحكم إحصائيا. وبالتالي، يتم تعريف معيار (*norm*) البيانات والشذوذ إحصائيا، اعتمادًا على التوزيع الاحتمالي للبيانات وتقدير مَعلَمات التوزيع.

تُعتبَر مخططات التحكم لشوارتز حساسةً للافتراض أن المتغير المستهدف يتبع توزيعًا طبيعيًا. أي انحراف عن هذا الافتراض الطبيعي قد يتسبب في أن يكون أداء مخطط التحكم لشوارتز، مثل مخطط التحكم لـــ $\bar{x}$، ضعيفًا، على سبيل المثال، إعطاء إشارة خارج السيطرة عندما تكون العملية في الحقيقة تحت السيطرة أو عدم إعطاء إشارة عندما تكون العملية هي في الحقيقة خارج السيطرة. نظرًا لأن مخططات التحكم لشوارتز ترصد وتقيم عينة بيانات واحدة فقط أو ملحوظة بيانات مرصودة فردية واحدة في كل مرة، فإنَّ مخططات التحكم لشوارتز ليست فعّالة في اكتشاف التحولات الصغيرة (*small shifts*)، على سبيل المثال، التحولات الصغيرة لمتوسط عملية ما والمراقبة بواسطة مخطط التحكم لـــ $\bar{x}$. تُعدُ مخططات تحكم المجموع التراكمي *CUSUM* في الجزء ١٦- ٢ ومخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA* في الجزء ١٦- ٣ أقل حساسيةً لافتراض طبيعية البيانات وهي فعّالة في اكتشاف التحولات الصغيرة. يمكن استخدام مخططات تحكم المجموع التراكمي *CUSUM* ومخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA* لمراقبة كلٍّ من عينات البيانات وملحوظات البيانات المرصودة الفردية. وبالتالي، تكون مخططات تحكم المجموع التراكمي *CUSUM* ومخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA* عمليةً أكثر.

١٦-٢ مخططات تحكم المجموع التراكمي (CUSUM Control Charts)

إذا كان لدينا سلسلة زمنية من ملحوظات البيانات المرصودة لمتغير x بحيث تكون الملحوظات المرصودة: $x_1, \ldots, x_n$ ، فإن المجموع التراكمي وصولًا إلى الملحوظة المرصودة رقم i هو (*Montgomery, 2001; Ye, 2003, Chapter 3*):

$$CS_i = \sum_{j=1}^{i} (x_i - \mu_0), \tag{١٦-٩}$$

حيث μ_0 هي القيمة الهدف لمتوسط العملية. إذا كانت العملية تحت السيطرة، فمن المتوقع أن تتذبذب ملحوظات البيانات المرصودة بشكل عشوائي حول متوسط العملية، وبالتالي يبقى CS_i حول الصفر. لكن، إذا كانت العملية خارجة عن السيطرة مع تحول لقيم x من متوسط العملية، فإن CS_i تظل في ازدياد إلى تحول موجب (أي، $x_i - \mu_0 > 0$) أو تظل في نقصان إلى تحول سالب. حتى إذا كان هناك تحول صغير، فإن أثر التحول الصغير يستمر بالتراكم في CS_i ويصبح كبيرًا إلى أن يتم اكتشاف خلله. وبالتالي، فإن مخطط تحكم المجموع التراكمي *CUSUM* يعدُّ أكثر فعّالية من مخطط التحكم لشوارتز للتحكم لاكتشاف التحولات الصغيرة لأن مخطط التحكم لشوارتز يفحص فقط عينة بيانات واحدة أو ملحوظة بيانات مرصودة واحدة. تُستخدم الصيغة ١٦-٩ لمراقبة ملحوظات البيانات المرصودة الفردية. إذا كان هناك امكانية لرصد عينات من سجلات البيانات، فإنه يمكن استبدال x_i في الصيغة ١٦-٩ بـ $\bar{x}_i$ لمراقبة متوسط العينة.

إذا كنا مهتمين باكتشاف تحول موجب فقط، فيمكن بناء مخطط تحكم المجموع التراكمي *CUSUM* من جانب واحد لمراقبة إحصائية CS_i^+:

$$CS_i^+ = \max[0, x_i - (\mu_0 + K) + CS_{i-1}^+], \tag{١٦-١٠}$$

حيث تُسمّى K القيمــة المرجعية التي تحدد مقدار الزيادة من متوسط العملية μ_0 الذي نحن مهتمون باكتشافه. ولأننا نتوقع أن تكــون $x_i \geq \mu_0 + K$ هي نتيجـــة لهـذا التحـــول الإيجــابــي K من متوســط العمليـــة μ_0، فنحـن نتوقـع أن تكـون $x_i - (K + \mu_0)$ موجبة ونتوقع أن تستمر CS_i^+ في الزيادة مع i. في حال أن بعض قيم x_i تجعل CS_{i-1}^++

$x_i - (\mu_0 + K)$ قيمة سالبة، فإن CS_i^+ تأخذ القيمة صفر وفقًا للصيغة ١٦-١٠ لأننا مهتمون فقط بالتحول الموجب. إحدى الطرق لتحديد قيمة K هي باستخدام معيار الانحراف σ من العملية. على سبيل المثال، $K = 0.5\sigma$ يشير إلى أننا مهتمون باكتشاف تحول 0.5σ فوق المتوسط المستهدف. إذا كانت العملية تحت السيطرة، فنحن نتوقع أن تبقى CS_i^+ حول الصفر. وبالتالي، يتم بدايةً تحديد قيمة CS_i^+ بالقيمة صفر:

$$CS_0^+ = 0. \qquad (١٦-١١)$$

عندما يتجاوز CS_i^+ حد القرار H، تُعتبَر العملية خارجة عن السيطرة. وعادةً ما تُستخدم $H = 5\sigma$ باعتبارها حد القرار بحيث يمكن تحقيق معدل منخفض للإنذارات الخاطئة (*Montgomery, 2001*). لاحظ أن $H = 5\sigma$ أكبر من حدود التحكم ٣-سيغما المستخدمة لمخطط التحكم $\bar{x}$ في الجزء ١٦-١ لأن CS_i^+ تُراكم تأثيرات ملحوظات البيانات المرصودة المتعددة بينما يقوم مخطط التحكم $\bar{x}$ بفحص ملحوظة بيانات واحدة أو عينة بيانات واحدة فقط.

إذا كنا مهتمين فقط باكتشاف تحولٍ سالبٍ، $-K$، من متوسط العملية، فإنه يمكن بناء مخطط تحكم المجموع التراكمي *CUSUM* بجانب واحد لمراقبة إحصائية CS_i^-:

$$CS_i^- = \max[0, (\mu_0 - K) - x_i + CS_{i-1}^-]. \qquad (١٦-١٢)$$

وحيث إننا نتـــوقع أن تكون $x_i \le \mu_0 - K$ نتيجـــةً للتحـول السـالب، $-K$، من متوســط العمليـة μ_0، فنتـــوقع أن تكـون $(\mu_0 - K) - x_i$ موجبة، ونتوقع أن تحافظ CS_i^- على الزياده مع i. وعادة ما تُستخدَم $H = 5\sigma$ باعتبارها حد القرار لتحقيق معدل منخفض للإنذارات الخاطئة (Montgomery, 2001). يتم بدايةً تحديد قيمة CS_i^- بالقيمة صفر لأننا نتوقع أن تظل CS_i^- قريبة من الصفر إذا كانت العملية تحت السيطرة:

$$CS_0^- = 0. \qquad (١٦-١٣)$$

يمكن استخدام مخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب مراقبة كلٍّ من: CS_i^+ باستخدام مخطط تحكم المجموع التراكمي *CUSUM* العلوي أحادي الجانب

و CS_i^- باستخدام مخطط تحكم المجموع التراكمي *CUSUM* السفلي أحادي الجانب لنفس x_i. إذا تجاوزت أيّ من CS_i^+ أو CS_i^- حدَ القرار H، تُعتبَر العملية خارجة عن السيطرة.

المثال ١٦-١

بالنظر إلى بيانات درجة حرارة الإطلاق (*Launch Temperature*) في الجدول ١-٥ والواردة في الجدول ١٦-٢ كسلسلة من ملحوظات البيانات المرصودة مع مرور الزمن. إذا كان لدينا المعلومات التالية:

$$\mu_0 = 69$$
$$\sigma = 7$$
$$K = 0.5\sigma = (0.5)(7) = 3.5$$
$$H = 5\sigma = (5)(7) = 35,$$

قم باستخدام مخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب لمراقبة درجة حرارة الإطلاق.

الجدول (١٦-٢)

ملحوظات البيانات المرصودة لدرجة حرارة الإطلاق من مجموعة بيانات الحلقات الدائرية ذات الأحمال الثقيلة جنباً إلى جنب مع الإحصائيات لمخطط تحكم المجموع التراكمي $CUSUM$ ثنائي الجانب

CS_i^-	CS_i^+	درجة حرارة الاطلاق x_i Launch Temperature x_i	ملحوظة البيانات المرصودة i Data Observation i
0	0	66	1
0	1	70	2
0	0	69	3
0	0	68	4
0	0	67	5
0	0	72	6
0	0.5	73	7
0	0	70	8
8.5	1	57	9
11	1	63	10
6.5	1	70	11
0	5.5	78	12
0	0	67	13
12.5	2	53	14
11	0	67	15
1.5	2.5	75	16
0	0	70	17
0	8.5	81	18
0	12	76	19
0	18.5	79	20
0	21	75	21
0	24.5	76	22
7.5	10	58	23

وبتحديد قيمة أولية لكل من CS_i^- وCS_i^+ مساويةً للصفر، مما يعني أنَّ $CS_0^+=0$ و$CS_0^-=0$. نقوم بحساب كلًا من CS_1^- وCS_1^+:

$$CS_1^+ = \max[0, x_1 - (\mu_0 + K) + CS_0^+] = \max[0, 66 - (69 + 3.5) + 0] = \max[0, -6.5] = 0$$

$$CS_1^- = \max[0, (\mu_0 - K) - x_1 + CS_0^-] = \max[0, (69 - 3.5) - 66 + 0] = \max[0, -0.5] = 0,$$

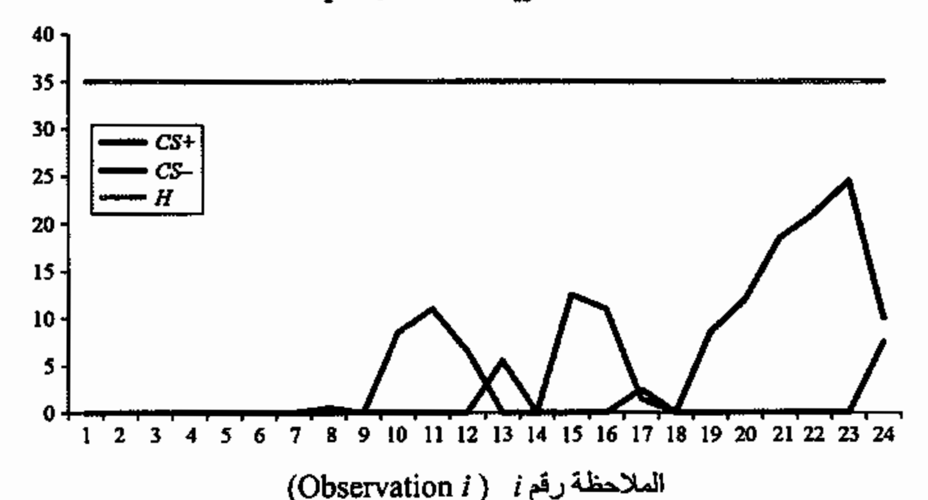
وبعدها نقوم بحساب CS_2^- وCS_2^+:

$$CS_2^+ = \max[0, x_2 - (\mu_0 + K) + CS_1^+] = \max[0, 70 - (69 + 3.5) + 0] = \max[0, -2.5] = 0$$

$$CS_2^- = \max[0, (\mu_0 - K) - x_2 + CS_1^-] = \max[0, (69 - 3.5) - 70 + 0] = \max[0, -4.5] = 0.$$

الشكل ١٦-١

مخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب لدرجة حرارة الإطلاق في مجموعة بيانات الحلقة الدائرية ذات الأحمال الثقيلة



وترد قيم CS_i^-و CS_i^+ لكل $i=3,...,23$ في الجدول ١٦-٢. يُظهر الشكل ١٦-١ مخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب. لا تتجاوز قيم CS_i^-و CS_i^+ لجميع الملحوظات المرصودة الـ ٢٣ حدّ القرار $H=35$. وبالتالي، لم يتم اكتشاف أي قيمة شاذة لدرجة حرارة الإطلاق. إذا تمّ تعيين حد القرار إلى $H=3\sigma=(3)(7)=21$، فسيتم الإشارة إلى الملحوظة المرصودة $i=22$ باعتبارها شاذة نظراً لأن $CS_{22}^+ = 24.5 > H$.

بعد أن يتم توليد إشارة خارج السيطرة، سوف يقوم مخطط تحكم المجموع التراكمي *CUSUM* بإعادة تهيئة CS_i^-و CS_i^+ إلى قيمهما الأولية الصفر، واستخدام القيمة الأولية المساوية للصفر لحساب CS_i^-و CS_i^+ للملحوظة التالية.

٣-١٦ مخططات التحكم للمتوسط المتحرك الموزون الأسّي (EWMA ControlCharts):

يعمل مخطط التحكم للمتوسط المتحرك الموزون الأسي *EWMA* لمتغير x وبملحوظات بيانات مرصودة مستقلة، x_i على مراقبة الإحصائية التالية (*Montgomery, 2001; Ye, 2003, Chapter 4*):

$$z_i = \lambda x_i + (1-\lambda) z_{i-1} \quad (١٤-١٦)$$

حيث λ عبارة عن وزن في النطاق *[0.1]*:

$$z_0 = \mu. \quad (١٥-١٦)$$

حدود التحكم هي (*Montgomery, 2001; Ye, 2003, Chapter 3*):

$$UCL = \mu + L\sigma\sqrt{\frac{\lambda}{2-\lambda}} \quad (١٦-١٦)$$

$$LCL = \mu - L\sigma\sqrt{\frac{\lambda}{2-\lambda}}. \quad (١٧-١٦)$$

يقوم الوزن λ بتحديد التأثيرات النسبية لملحوظة البيانات المرصودة الحالية، x_i وملحوظات البيانات المرصودة السابقة كما تمّ التقاطها من خلال z_{i-1} على z_i. إذا عبرنا عن z_i باستخدام x_i حيث $x_{i-1}, \ldots, x_1$:

$$\begin{aligned} z_i &= \lambda x_i + (1-\lambda) z_{i-1} \\ &= \lambda x_i + (1-\lambda)[\lambda x_{i-1} + (1-\lambda) z_{i-2}] \\ &= \lambda x_i + (1-\lambda)\lambda x_{i-1} + (1-\lambda)^2 z_{i-2} \\ &= \lambda x_i + (1-\lambda)\lambda x_{i-1} + (1-\lambda)^2[\lambda x_{i-2} + (1-\lambda) z_{i-3}] \\ &= \lambda x_i + (1-\lambda)\lambda x_{i-1} + (1-\lambda)^2 \lambda x_{i-2} + (1-\lambda)^3 z_{i-3} \\ &\dots \\ &= \lambda x_i + (1-\lambda)\lambda x_{i-1} + (1-\lambda)^2 \lambda x_{i-2} + \dots + (1-\lambda)^{i-2}\lambda x_2 + (1-\lambda)^{i-1}\lambda x_1 \end{aligned} \quad (١٦-١٨)$$

يمكننا ملاحظة أن الأوزان x_i حيث $x_{i-1},\dots,x_1$ ، تتناقص بشكل أسّي، فعلى سبيل المثال، عندما تكون $\lambda=0.3$ يكون الوزن 0.3 لـ x_i و$0.21 = (0.3)\ (0.7)$ لـ x_{i-1} و$0.147= (0.3)^2\ (0.7)$ لـ x_{i-2}، و$0.1029= (0.3)^3\ (0.7)$ لـ x_{i-3} ...، كما هو موضح في الشكل ١٦-٢. وهذا المصطلح يُسمَى مخطط تحكم المتوسط المتحرك الموزون الأسي $EWMA$. كلما كانت قيمة λ أكبر كان تأثير ملحوظات البيانات المرصودة السابقة أقل، وكان تأثير ملحوظة البيانات المرصودة الحالية أكثر على إحصائية $EWMA$ الحالية، z_i.

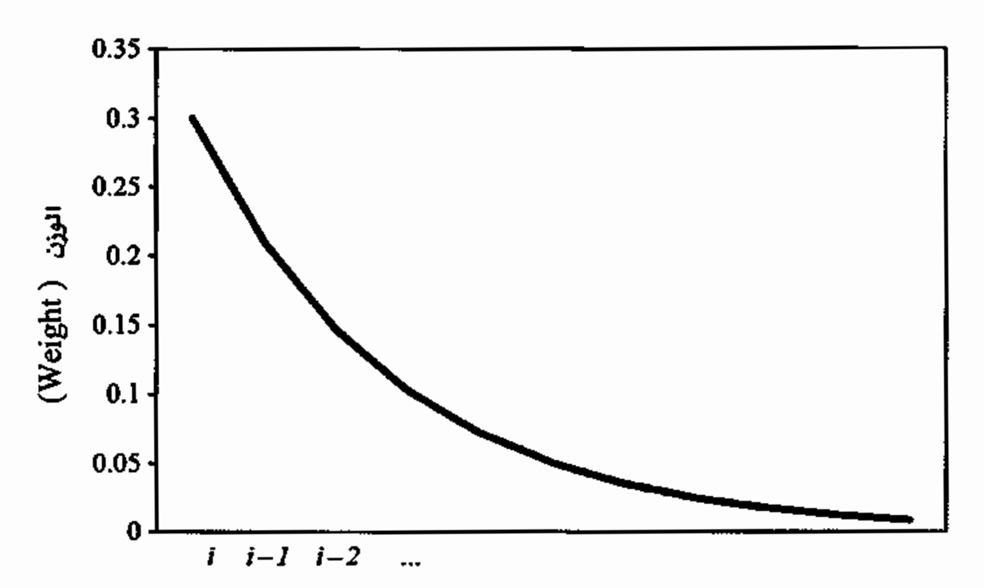
في المعادلات من ١٦-١٤ وحتى ١٦-١٧، عادةً ما يعمل إسناد قيم لـــ λ وL ضمن النطاقات التالية بشكل جيد (*Montgomery, 2001; Ye, 2003, Chapter 4*):

$$0.05 \le \lambda \le 0.25$$
$$2.6 \le L \le 3.$$

ويمكن استخدام عينة بيانات لحساب متوسط العينة والانحراف المعياري للعينة كتقديرات لكل من μ وσ، على التوالي.

الشكل (١٦-١)

أوزان متناقصة أسياً على ملحوظات البيانات المرصودة



المثال ١٦-٢

بالنظر إلى بيانات درجة حرارة الإطلاق (*Launch Temperature*) في الجدول ١-٥ والواردة في الجدول ١٦-٣ كسلسلة من ملحوظات البيانات المرصودة مع مرور الوقت. إذا كان لدينا ما يلي:

$$\mu = 69$$
$$\sigma = 7$$
$$\lambda = 0.2$$
$$L = 3,$$

قم باستخدام مخطط تحكم المتوسط المتحرك الموزون الأسّي *EWMA* لمراقبة درجات حرارة الإطلاق.

علينا أولًا أن نحسب حدود التحكم (*control limits*):

$$UCL = \mu + L\sigma\sqrt{\frac{\lambda}{2-\lambda}} = 69 + (3)(7)\sqrt{\frac{0.3}{2-0.3}} = 77.82$$

$$LCL = \mu - L\sigma\sqrt{\frac{\lambda}{2-\lambda}} = 69 - (3)(7)\sqrt{\frac{0.3}{2-0.3}} = 60.18.$$

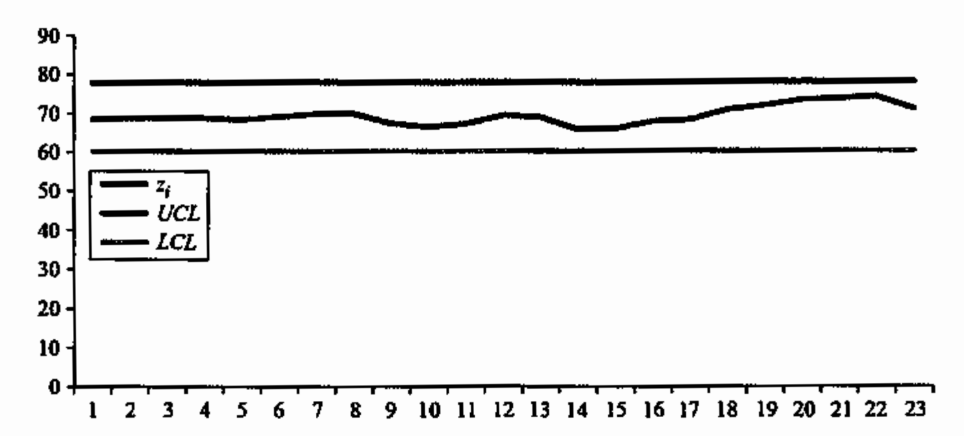
الجدول (١٦-٣)

ملحوظات البيانات المرصودة لدرجة حرارة الإطلاق مجموعة بيانات الحلقات الدائرية ذات الأحمال الثقيلة جنباً إلى جنب مع إحصائية *EWMA* لمخطط تحكم الـ *EWMA*

z_i	درجة حرارة الاطلاق x_i **Launch Temperature x_i**	ملحوظة البيانات المرصودة i **Data Observation i**
68.4	66	1
68.72	70	2
68.78	69	3
68.62	68	4
68.30	67	5
69.04	72	6
69.83	73	7
69.86	70	8
67.29	57	9
66.43	63	10
67.15	70	11
69.32	78	12
68.85	67	13
65.68	53	14
65.95	67	15
67.76	75	16
68.21	70	17
70.76	81	18
71.81	76	19
73.25	79	20
73.60	75	21
74.08	76	22
70.86	58	23

الشكل (١٦-٣)

مخطط تحكم *EWMA* لمراقبة درجة حرارة الإطلاق من مجموعة بيانات الحلقات الدائرية ذات الأحمال الثقيلة



الملحوظة المرصودة رقم *i* (Observation *i*)

باستخدام $z_0=\mu=69$ ، نقوم بحساب إحصائية *EWMA*:

$$z_1 = \lambda x_1 + (1-\lambda)z_0 = (0.2)(66) + (1-0.2)(69) = 68.4$$

$$z_2 = \lambda x_2 + (1-\lambda)z_1 = (0.2)(70) + (1-0.2)(68.4) = 68.72$$

وترد قيم إحصائية *EWMA* لملحوظات البيانات المرصودة الأخرى في الجدول ١٦-٣. تبقى قيم إحصائية *EWMA* لجميع ملحوظات البيانات المرصودة الـ ٢٣ ضمن حدود التحكم، [*LCL* , *UCL*] = [*60.18, 77.82*] ، ولا يتم اكتشاف أي قيم شاذة. يعرض الشكل ١٦-٣ مخطط تحكم *EWMA* مع إحصائية *EWMA* وحدود التحكم.

إذا تمّ ربط ملحوظات البيانات المرصودة ذاتيًا (انظر الفصل ١٨ لشرح الارتباط الذاتي*autocorrelation*)، فانه يمكننا أولًا بناء نموذج التنبؤ بخطوة واحدة للأمام (*1-step ahead prediction model*) من البيانات المرتبطة ذاتيا، ومقارنة ملحوظة بيانات مرصودة معينة مع قيمتها التنبئية بخطوة واحدة للأمام من أجل الحصول على الخطأ

(*error*) أو المتبقي (*residual*)، واستخدام مخطط تحكم الـ *EWMA* لرصد البيانات المتبقية (*residual data*) (*Montgomery and Mastrangelo, 1991*). يتم حساب قيمة التنبؤ بخطوة واحدة للأمام لـ x_i على النحو التالي:

$$z_{i-1} = \lambda x_{i-1} + (1-\lambda) z_{i-2} \,, \qquad (١٩-١٦)$$

حيث $0 < \lambda \leq 1$ وهذا يعني أن z_{i-1} هو المتوسط المتحرك الموزون الأسي *EWMA* لـ $x_{i-1}, \ldots, x_1$ ويُستخدَم كتنبؤ لـ x_i. ثم يتم احتساب خطأ التنبؤ أو المتبقي كما يلي:

$$e_i = x_i - z_{i-1} \,. \qquad (٢٠-١٦)$$

في المعادلة ١٦-١٩، يمكن تعيين λ لتخفيض مجموع أخطاء التنبؤ التربيعية على مجموعة البيانات الاستكشافية أو التدريبية:

$$\lambda = \arg\min_{\lambda} \sum_i e_i^2 \,. \qquad (٢١-١٦)$$

إذا كان نموذج التنبؤ بخطوة واحدة للأمام يمثل البيانات المترابطة ذاتيًا بشكل جيد، ينبغي أن تكون قيم $e_i s$ مستقلةً عن بعضها وتكون موزعةً طبيعيًا بمتوسط يساوي صفر وانحراف معياري يساوي σ_e. يكون محور مخطط تحكم المتوسط المتحرك الموزون الأسي *EWMA* لمراقبة e_i عند مستوى الصفر كما أن لديه حدود التحكم التالية:

$$UCL_{e_i} = L\hat{\sigma}_{e_i-1} \qquad (٢٢-١٦)$$

$$LCL_{e_i} = -L\hat{\sigma}_{e_i-1} \qquad (٢٣-١٦)$$

$$\hat{\sigma}^2_{e_i-1} = \alpha e^2_{i-1} + (1-\alpha)\hat{\sigma}^2_{e_i-2}, \qquad (٢٤-١٦)$$

حيث يتم تحديد L إلى قيمة بحيث $2.6 \leq L \leq 3$، و$0 < \alpha \leq 1$ وتعطي $\hat{\sigma}^2_{e_{i-1}}$ تقدير القيمة σ_e لـ x_i باستخدام المتوسط المتحرك الموزون الأسي *EWMA* لأخطاء التنبؤ. باستخدام المعادلة ١٦-٢٠، والتي تعطي $x_i = e_i + z_{i-1}$ فإنّ التحكم لرصد x_i مباشرةً بدلًا من e_i هو:

$$UCL_{x_i} = z_{i-1} + L\hat{\sigma}_{e_i-1} \tag{١٦-٢٥}$$

$$LCL_{x_i} = z_{i-1} - L\hat{\sigma}_{e_i-1} \tag{١٦-٢٦}$$

على غرار مخطط تحكم المجموع التراكمي *CUSUM*، يُعتَبر مخطط تحكم المتوسط المتحرك الموزون الأسي *EWMA* أكثر صلابةً لفرضية طبيعية توزيع البيانات من مخططات التحكم لشوارتز (*Montgomery, 2001*). خلافًا لمخططات التحكم لشوارتز، فإن مخططات تحكم المجموع التراكمي *CUSUM* ومخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA* تُعتبَر فعّالة في اكتشاف الحالات الشاذة ليس فقط للتحولات الكبيرة ولكن أيضًا للتحولات الصغيرة لأن مخططات تحكم المجموع التراكمي *CUSUM* ومخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA* تأخذ في الاعتبار التأثيرات الخاصة بملحوظات البيانات المرصودة المتعددة.

١٦-٤ مخططات تحكم الدرجة التراكمية (Cuscore Contol Charts) :

تكشف مخططات التحكم الموصوفة في الأجزاء من ١٦-١ وحتى ١٦-٣ عن التحولات الخارجة عن السيطرة من المتوسط أو الانحراف المعياري. تمّ تصميم مخططات تحكم الدرجة التراكمية (*cuscore*) (*Luceno, 1999*) للكشف عن التغير من أي شكل معين من نموذج بيانات تحت السيطرة إلى أي شكل معين من نموذج بيانات خارج السيطرة. على سبيل المثال، يمكن بناء مخطط تحكم الدرجة التراكمية (*Cuscore*) للكشف عن تغير الميل في نموذج خطي لبيانات تحت السيطرة على النحو التالي:

نموذج بيانات تحت السيطرة:

$$y_t = \theta_0 t + \varepsilon_t \qquad (١٦-٢٧)$$

نموذج بيانات خارج السيطرة:

$$y_t = \theta t + \varepsilon_t\,, \qquad \theta \neq \theta_0\,, \qquad (١٦-٢٨)$$

حيث إن ε_t هو متغير عشوائي بتوزيع طبيعي، والمتوسط $\mu = 0$ والانحراف المعياري σ. وبمثال آخر، يمكن أن يكون لدينا مخطط تحكم درجة تراكمية لاكتشاف وجود موجة جيبية (*Sine wave*) داخل عملية تحت السيطرة مع وجود تباينات عشوائية من المتوسط T:

نموذج بيانات تحت السيطرة:

$$y_t = T + \theta_0 \sin\left(\frac{2\pi t}{p}\right) + \varepsilon_t\,, \qquad \theta_0 = 0\,, \qquad (١٦-٢٩)$$

نموذج بيانات خارج السيطرة:

$$y_t = T + \theta \sin\left(\frac{2\pi t}{p}\right) + \varepsilon_t. \qquad (١٦-٣٠)$$

لبناء إحصائية الدرجة التراكمية *Cuscore*، نأخذ في الاعتبار قيمة y_t كدالة عن x_t والمعلمة θ والتي تميز عملية خارج السيطرة عن عملية تحت السيطرة:

$$y_t = f(x_t, \theta) \qquad (١٦-٣١)$$

وعندما تكون العملية تحت السيطرة، يكون لدينا:

$$\theta = \theta_0. \qquad (١٦-٣٢)$$

في المثالين الموضحين في المعادلات من ١٦-٢٧ وحتى ١٦-٣٠، فإن x_t تحتوي t فقط، وθ_0 = θ عندما تكون العملية تحت السيطرة.

يمكن حساب المتبقي، ε_t، عن طريق طرح القيمة المتوقعة $\hat{y}_t$ من القيمة المرصودة y_t:

$$\varepsilon_t = y_t - \hat{y}_t = y_t - f(x_t, \theta) = g(y_t, x_t, \theta). \qquad (١٦\text{-}٣٣)$$

عندما تكون العملية تحت السيطرة، يصبح لدينا $\theta=\theta_0$ ونتوقع أن تكون $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ مستقلةً عن بعضها البعض، وكل منها عبارة عن متغير عشوائي غير مرتبط بمتغيرات عشوائية أخرى مع ملحوظات بيانات مرصودة ومستقلة، وبتوزيع طبيعي، وبمتوسط $\mu= 0$ وبانحراف معياري σ. وهذا يعني أن المتغيرات العشوائية، $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ لها توزيع طبيعي مشترك متعدد المتغيرات وبدالة الكثافة الاحتمالية المشتركة التالية:

$$P(\varepsilon_1, ..., \varepsilon_n | \theta = \theta_0) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{t=1}^{n}\frac{\varepsilon_{t0}^2}{\sigma^2}}. \qquad (١٦\text{-}٣٤)$$

وبأخذ اللوغاريتم الطبيعي للمعادلة ١٦-٣٤، يصبح لدينا:

$$l(\varepsilon_1, ..., \varepsilon_n | \theta = \theta_0) = -\frac{2}{n} ln(2\pi) - \frac{1}{2\sigma^2}\sum_{t=1}^{n} \varepsilon_{t0}^2. \qquad (١٦\text{-}٣٥)$$

كما يتضح من المعادلة ١٦-٣٣، فإن ε_t هي دالة من θ، $P\ (\varepsilon_1, ..., \varepsilon_n)$ في المعادلة ١٦-٣٤ تصل إلى قيمة الإمكان القصوى (*maximum likelihood*) إذا كانت العملية تحت السيطرة مع $\theta = \theta_0$ يكون لدينا ε_{t0}، حيث $t=1, ..., n$ ، الموزعة بشكل طبيعي ومستقل ومتطابق، متواجدة في معادلة ١٦-٣٤. إذا كانت العملية خارج السيطرة وكانت $\theta \neq \theta_0$، فلا تكون المعادلة ١٦-٣٤ دالة كثافة الاحتمال المشترك الصحيحة لـ $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ وبالتالي لا تعطي قيمة الإمكان القصوى لـ $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$. وبالتالي، إذا كانت العملية تحت السيطرة مع $\theta = \theta_0$ ، يكون لدينا:

$$\frac{\partial l(\varepsilon_1, \dots, \varepsilon_n | \theta = \theta_0)}{\partial \theta} = 0. \qquad (١٦-٣٦)$$

باستخدام المعادلة ١٦-٣٥ للتعويض عن $l\ (\varepsilon_1, \dots, \varepsilon_n \mid \theta = \theta_0)$ في المعادلة ١٦-٣٦ وإسقاط جميع حدود المعادلة التي لا علاقة لها بـ θ عند عمل الاشتقاق، يصبح لدينا:

$$\sum_{t=1}^{n} \varepsilon_{t0} \left(-\frac{\partial \varepsilon_{t0}}{\partial \theta} \right) = 0. \qquad (١٦-٣٧)$$

تكون إحصائية الدرجة التراكمية *Cuscore* لمخطط تحكم الدرجة التراكمية للمراقبة مساوية لـــ Q_0:

$$Q_0 = \sum_{t=1}^{n} \varepsilon_{t0} \left(-\frac{\partial \varepsilon_{t0}}{\partial \theta} \right) = \sum_{t=1}^{n} \varepsilon_{t0}\, d_{t0} \qquad (١٦-٣٨)$$

حيث:

$$d_{t0} = -\frac{\partial \varepsilon_{t0}}{\partial \theta}. \qquad (١٦-٣٩)$$

وبناءً على المعادلة ١٦-٣٧، من المتوقع أن تظل Q_0 قريبةً من الصفر إذا كانت العملية تحت السيطرة مع $\theta = \theta_0$ إذا انحازت θ عن θ_0، فإن قيمة Q_0 تنحرف عن منطقة الصفر بطريقة ليست عشوائية، بل بطريقة متسقة.

على سبيل المثال، لاكتشاف أي تغير على ميل نموذج خطي لبيانات تحت السيطرة الموضحة في المعادلات ١٦-٢٧ و١٦-٢٨، فإن مخطط تحكم الدرجة التراكمية *Cuscore* يقوم بمراقبة القيمة:

$$Q_0 = \sum_{t=1}^{n} \varepsilon_{t0}\left(-\frac{\partial \varepsilon_{t0}}{\partial \theta}\right) = \sum_{t=1}^{n} \varepsilon_{t0}\left(-\frac{\partial (y_t - \theta t)}{\partial \theta}\right) = \sum_{t=1}^{n} (y_t - \theta_0 t)t. \qquad (٤٠-١٦)$$

إذا كان الميل θ للنموذج الخطي تحت السيطرة الذي يتغير من θ_0 فإن $(y_t - \theta_0 t)$ في المعادلة ١٦- ٤٠ يحتوي على t، الذي يتم ضربه في قيمة أخرى لـ t لجعل Q_0 يستمر في الزيادة (إذا $y_t - \theta_0 t > 0$) أو في النقصان (إذا $y_t - \theta_0 t < 0$) بدلًا من التغيير عشوائيا قريبًا من الصفر. هذا الانطلاق المستمر لقيم Q_0 من الصفر يتسبب في أن يزيد أو ينقص ميل الخط، الذي يربط قيم Q_0 مع مرور الوقت، من الصفر، الأمر الذي يمكن استخدامه كإشارة إلى وجود حالة شاذة.

لاكتشاف موجة جيبية في عملية تحت السيطرة بمتوسط T تباينات عشوائية مبينة في المعادلات ١٦-٢٩ و١٦-٣٠، تكون إحصائية الدرجة التراكمية *Cuscore* لمخطط تحكم الدرجة التراكمية هي:

$$Q_0 = \sum_{t=1}^{n} \varepsilon_{t0}\left(-\frac{\partial \varepsilon_{t0}}{\partial \theta}\right) = \sum_{t=1}^{n} (y_t - T)\left[-\frac{\partial \left(y_t - T - \theta \sin\left(\frac{2\pi t}{p}\right)\right)}{\partial \theta}\right]$$

$$= \sum_{t=1}^{n} (y_t - T)\sin\left(\frac{2\pi t}{p}\right). \qquad (٤١-١٦)$$

إذا كانت الموجة الجيبية موجودة في y_t، فإن $(y_t - T)$ في المعادلة ١٦- ٤١ تحتوي على $sin(2\pi t/p)$، والتي يتم ضربها في قيمة أخرى لـ $sin(2\pi t/p)$ لجعل Q_0 تستمر في الزيادة (إذا كانت $y_t - T > 0$) أو في النقصان (إذا كانت $y_t - T < 0$) بدلًا من التغيير عشوائيًا حول الصفر.

لاكتشاف تحول المتوسط K من μ_0 كما في مخطط تحكم المجموع التراكمي *CUSUM* الموضح في المعادلات ١٦- ٩، ١٦- ١٠، و١٦- ١٢، يكون لدينا:

نموذج البيانات تحت السيطرة:

$$y_t = \mu_0 + \theta_0 K + \varepsilon_t\,, \quad \theta_0 = 0 \qquad (١٦-٤٢)$$

نموذج البيانات خارج السيطرة:

$$y_t = \mu_0 + \theta K + \varepsilon_t\,, \quad \theta \neq \theta_0 \qquad (١٦-٤٣)$$

$$Q_0 = \sum_{t=1}^{n} \varepsilon_{t0}\left(-\frac{\partial \varepsilon_{t0}}{\partial \theta}\right) = \sum_{t=1}^{n}(y_t - \mu_0)\left[-\frac{\partial(y_t - \mu_0 - \theta K)}{\partial \theta}\right] = \sum_{t=1}^{n}(y_t - \mu_0)K. \qquad (١٦-٤٤)$$

في حالة حدوث تحول المتوسط K من μ_0، فإن $(y_t - \mu_0)$ في المعادلة ١٦- ٤٤ يحتوي على K، والذي يكون مضروباً في قيمة أخرى لـ K لجعل Q_0 يستمر في الزيادة (إذا كانت $y_t - \mu_0 > 0$) أو في النقصان (إذا كانت $y_t - \mu_0 < 0$) بدلًا من التغير عشوائيا حول الصفر.

حيث إن مخططات تحكم الدرجة التراكمية *Cuscore* تسمح لنا باكتشاف نموذج معين لحالة شاذة إذا كان معطى لنا نموذجًا معيناً لنموذج بيانات تحت السيطرة، فإن مخططات تحكم الدرجة التراكمية تسمح لنا برصد واكتشاف مجموعة واسعة من حالات تحت السيطرة مقابل حالات خارج السيطرة بشكل أكثر من مخططات التحكم لشوارتز، ومخططات تحكم المجموع التراكمي *CUSUM*، ومخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA*.

١٦-٥ منحنى التشغيل التشخيصي لتقييم ومقارنة مخططات التحكم:

(Receiver Operating Curve –ROC- for Evaluation and Comparison of Control Charts)

تنتج القيم المختلفة لمعلَمات حد القرار والمستخدمة في مخططات تحكم متنوعة، على سبيل المثال، ٣- سيغما في مخطط تحكم $\bar{x}$، وH في مخطط تحكم المجموع التراكمي *CUSUM*، وL في مخطط تحكم المتوسط المتحرك الموزون الأسي *EWMA* معـدلات

مختلفـــةً من الإنذارات الخاطئة والزيارات الناجحة. لنفتـــرض في المثـال ١٦- ١ أن أي قيمة لـ $x_i \geq 75$ هي في الحقيقة حالة شاذة. وبالتالي، يكون لدينا ملحوظات البيانات المرصودة السبع، وهي الملحوظات أرقام ١٢، ١٦، ١٨، ١٩، ٢٠، ٢١، و٢٢، لديها $x_i \geq 75$ وهي بالفعل حالات شاذة. إذا تمّ تعديل قيمة حد القرار إلى قيمة أكبر من أو يساوي الحد الأقصى لقيمة CS_i^-و CS_i^+ لجميع ملحوظات البيانات المرصودة الـ ٢٣، على سبيل المثال، $H=24.5$، فإن CS_i^-و CS_i^+ لجميع ملحوظات البيانات المرصودة الـ ٢٣ لا تتجاوز H، ومخطط تحكم المجموع التراكمي $CUSUM$ ثنائي الجانب لا يعطي إشارة إلى أي ملحوظة بيانات مرصودة باعتبارها ملحوظة شاذة. ولا يكون لدينا أي إنذارات خاطئة كما أن عدد الزيارات الناجحة صفر، وهذا يعني، أن لدينا معدل الإنذار الخاطئ ٠٪ ومعدل الزيارة الناجحة ٠٪. إذا تمّ تعديل قيمة حد القرار إلى قيمة أصغر من قيمة الحد الأدنى لقيمة CS_i^-و CS_i^+ لجميع ملحوظات البيانات المرصودة الـ ٢٣، على سبيل المثال، $H= -1$، فإن CS_i^-و CS_i^+ لجميع ملحوظات البيانات المرصودة الـ ٢٣ تتجاوز H، ويقوم مخطط تحكم المجموع التراكمي $CUSUM$ ثنائي الجانب بعمل إشارة إلى كل ملحوظة بيانات مرصودة على أنها ملحوظة شاذة، مما ينتج ٧ زيارات ناجحة على جميع الملحوظات الشاذة الصحيحة (الملحوظات هي أرقام ١٢، ١٦، ١٨، ١٩، ٢٠، ٢١، و٢٢) و١٦ إنذارا خاطئًا، وهذا يعني أن معدل الزيارة الناجحة هو ١٠٠٪ ومعدل الإنذار الخاطئ هو ١٠٠٪. إذا تمّ تعديل قيمة حد القرار إلى $H=0$، فإن مخطط تحكم المجموع التراكمي $CUSUM$ ثنائي الجانب يعطي إشارة إلى ملحوظات البيانات المرصودة أرقام ٧، ٩، ١٠، ١١، ١٢، ١٤، ١٥، ١٦، ١٨، ١٩، ٢٠، ٢١، و٢٢ على أنها ملحوظات شاذة، مما ينتج إشارات خارج السيطرة عددها ٧ على كل الملحوظات السبع الشاذة الحقيقية (معدل الزيارة الناجحة١٠٠٪) و٧ إشارات خارج السيطرة على الملحوظات أرقام ٧، ٩، ١٠، ١١، ١٤، ١٥ و٢٣ من أصل ١٦ ملحوظة بيانات مرصودة تحت السيطرة (معدل إنذار خاطئ ٤٤٪). يسرد الجدول ١٦-٤ أزواج معدل الإنذار الخاطئ ومعدل الزيارة الناجحة لقيم أخرى لـ H لمخطط تحكم المجموع التراكمي $CUSUM$ ثنائي الجانب في المثال ١٦-١.

يعرض منحنى التشغيل التشخيصي (ROC) بيانيًا أزواجًا من معدل الزيارة الناجحة ومعدل الإنذار الخاطئ لقيم متنوعة من حد القرار. يعرض الشكل ١٦-٤ منحنى التشغيل التشخيصي (*Receiver Operating Curve-ROC*) لمخطط تحكم المجموع التراكمي

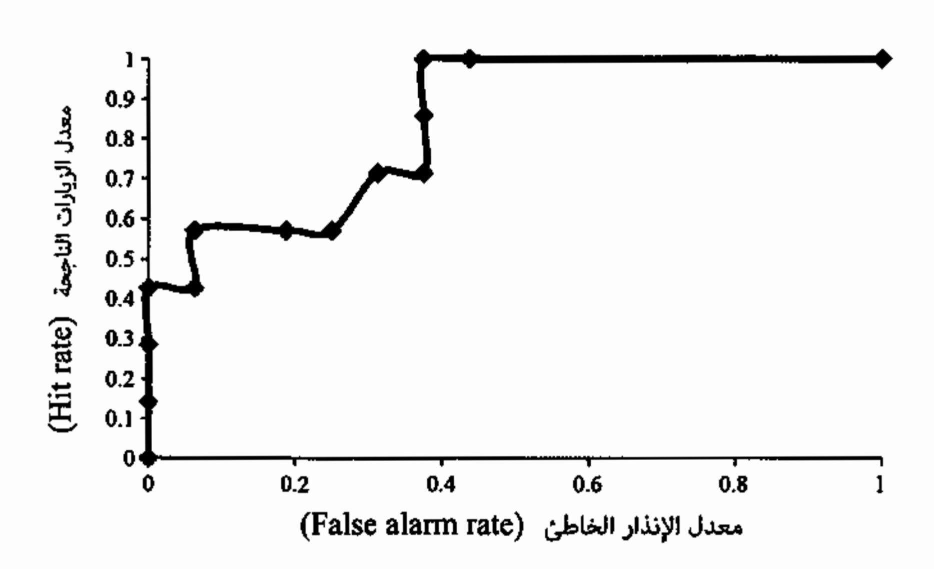
$CUSUM$ ثنائي الجانب في المثال ١٦-١، إذا كان معطى لدينا سبع حالات شاذة حقيقية على الملحوظات المرصودة أرقام ١٢، ١٦، ١٨، ١٩، ٢٠، ٢١، و٢٢، وعلى عكس زوج من معدل الإنذار الخاطئ ومعدل الزيارة الناجحة لقيمة معينة من حد القرار، فإن منحنى التشغيل التشخيصي (ROC) يعطي صورةً كاملةً عن الأداء من خلال تقنية اكتشاف الوضع الشاذ.

الجدول (١٦-٤)

أزواج من معدل الإنذار الخاطئ ومعدل الزيارة الناجحة لقيم متنوعة من حد القرار H لمخطط تحكم المجموع التراكمي $CUSUM$ ثنائي الجانب في المثال ١٦-١

معدل الزيارات الناجحة **Hit Rate**	معدل الإنذار الخاطئ **False Alarm Rate**	***H***
1	1	-1
1	0.44	0
1	0.38	0.5
0.86	0.38	2.5
0.71	0.38	5.5
0.71	0.31	6.5
0.57	0.25	8.5
0.57	0.19	10
0.57	0.06	11
0..43	0.06	12
0.43	0	12.5
0.29	0	18.5
0.14	0	21
0	0	24.5

الشكل (١٦-٤)
منحنى التشغيل التشخيصي (*ROC*) لمخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب في المثال ١٦-١



كلما اقترب منحنى التشغيل التشخيصي (*ROC*) من أعلى الزاوية اليسرى، التي تمثل معدل الإنذار الخاطئ (٠٪) ومعدل الزيارة الناجحة (١٠٠٪)، للمخطط، كلما كان الأداء أفضل لمخرجات تقنية اكتشاف الحالات الشاذة. ونظراً لأنه من الصعب تثبيت استخدام حدود القرار لتقنيتين مختلفتين لاكتشاف الحالات الشاذة بحيث يمكن مقارنة أدائهما بشكل عادل، فإن منحنى التشغيل التشخيصي (*ROC*) يمكن رسمه بيانياً لكل طريقة تقنية في نفس المخطط لمقارنة منحنيات التشغيل التشخيصية (*ROCs*) لتقنيتين اثنتين ودراسة أي منحنى تشغيل تشخيصي (*ROC*) يكون أقرب إلى الزاوية العلوية اليسرى للمخطط لتحديد أي تقنية تعطي أداء أفضل للاكتشاف. يوضح يي وآخرون *(Ye et al., 2002b)* استخدام منحنيات التشغيل التشخيصية (*ROCs*) لمقارنة أداء اكتشاف الهجوم الإلكتروني (عبر الإنترنت) باستخدام تقنتين اثنتين من مخططات التحكم.

١٦-٦ البرمجيات والتطبيقات (Software and Applications) :

يدعم برنامج *Minitab* *(www.minitab.com)* مخططات تحكم العملية الإحصائية. يمكن العثور على تطبيقات لمخططات التحكم أحادية المتغير لجودة التصنيع واكتشاف التسلل عبر الإنترنت في *(Ye, 2003, Chapter 3)*، *(Ye, 2008)*، *(Ye et al., 2002a, 2004)*، و*(Ye and Chen, 2003)*.

التمارين (Exercises)

١٦-١ بالنظر إلى بيانات درجة حرارة الإطلاق (*Launch Temperature*) والمعلومات التالية في المثال ١٦-١:

$$\mu_0 = 69$$
$$K = 3.5$$

قم ببناء مخطط تحكم الدرجة التراكمية *Cuscore* باستخدام المعادلة ١٦-٤٤ لمراقبة درجة حرارة الإطلاق.

١٦-٢ ارسم منحنيات التشغيل التشخيصية (*ROCs*) لمخطط تحكم المجموع التراكمي *CUSUM* في المثال ١٦-١، ومخطط تحكم المتوسط المتحرك الموزون الأسي *EWMA* في المثال ١٦-٢، ومخطط تحكم الدرجة التراكمية *Cuscore* في التمرين ١٦-١، في نفس المخطط، ومقارنة أداء تقنيات مخطط التحكم هذه.

١٦-٣ قم بجمع بيانات درجات الحرارة اليومية في الأشهر الـ ١٢ الأخيرة في مدينتك، واعتبر بيانات درجة الحرارة في كل شهر كعينة البيانات، وقم ببناء مخطط تحكم $\bar{x}$ لمراقبة درجات الحرارة المحلية واكتشاف أي حالات شاذة.

١٦-٤ بالنظـــر إلى مجموعـــة البيانات نفسها التي تتكون من ١٢ متوسط درجات حرارة شهـــرية التي تمّ الحصول عليها من التمـــرين ١٦-٣ وقـــم باستخدام $\bar{\bar{x}}$ و$\bar{s}$ التي تمّ الحصول عليـــها من التمـــرين ١٦-٣ لتقـــدير μ_0 وσ. قم بتعديل $K= 0.5\sigma$، و$H= 5\sigma$. قم ببناء مخطط تحكم المجموع التراكمي *CUSUM* ثنائي الجانب لمراقبة بيانات متوسط درجات الحرارة الشهرية واكتشاف أي حالات شاذة.

١٦-٥ بالنظر إلى مجموعة البيانات وقيم μ_0 وK في التمرين ١٦-٤. قم ببناء مخطط تحكم الدرجة التراكمية *Cuscore* لمراقبة بيانات متوسط درجات الحرارة الشهرية واكتشاف أي حالات شاذة.

١٦-٦ بالنظر إلى مجموعة البيانات وتقديرات كل من μ_0 وσ في التمرين ١٦-٤. قم بتحديد $\lambda = 0.1$ و$L = 3$ قم ببناء مخطط تحكم المتوسط المتحرك الموزون الأسي *EWMA* لمراقبة بيانات متوسط درجات الحرارة الشهرية.

١٦-٧ كرّر التمرين ١٦-٦ ولكن مع $\lambda = 0.3$ وقم بمقارنة مخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA* في التمارين ١٦-٦ و١٦-٧.

١٧- مخططات التحكم متعددة المتغيرات
Multivariate Control Charts

تعمل مخططات التحكم متعددة المتغيرات (*Multivariate control charts*) على مراقبة ورصد متغيرات متعددة في وقت واحد لاكتشاف الحالات الشاذة. يصف هذا الفصل ثلاثة من مخططات التحكم الإحصائية المتعددة المتغيرات، وهي: مخططات التحكم لهوتلينق (*Hotelling's T^2 control charts*)، ومخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA* متعددة المتغيرات (*multivariate EWMA control charts*)، ومخططات تحكم مربع كاي (*chi-square control charts*). كما سنتناول في هذا الفصل بعض التطبيقات الخاصة بمخططات التحكم متعددة المتغيرات مع المراجع.

١-١٧ مخططات التحكم T^2 لهوتلينق (Hotelling's T^2 Control Charts)

لنجعل $x_i=(x_{i1},...,x_{ip})$ ترمز إلى ملحوظة البيانات المرصودة رقم i للمتغيرات العشوائية، $x_{i1},...,x_{ip}$ التي تتبع توزيعًا طبيعيًا متعدد المتغيرات (انظر إلى دالة الكثافة الاحتمالية للتوزيع الطبيعي متعدد المتغيرات في الفصل ١٦) وبالمتجه المتوسط $\boldsymbol{\mu}$ ومصفوفة التباين- التغاير $\boldsymbol{\Sigma}$ (انظر إلى تعريف مصفوفة التباين- التغاير في الفصل ١٤). إذا كان لدينا عينة بيانات بعدد n من ملحوظات البيانات المرصودة، فإن المتجه المتوسط للعينة $\bar{\boldsymbol{x}}$ ومصفوفة التباين- التغاير للعينة $\boldsymbol{S}$:

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad (١-١٧)$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})', \qquad (٢-١٧)$$

يمكن استخدامها لتقدير قيمة كل من μ وΣ على التوالي. إحصاءة هوتلينق T^2 لملحوظة بيانات مرصودة، x_i هي (*Chou et al., 1999; Everitt, 1979; Johnson and Whichern, 1998; Mason et al., 1995, 1997a,b; Mason and Young, 1999; Ryan, 1989*):

$$T^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}), \qquad (١٧-٣)$$

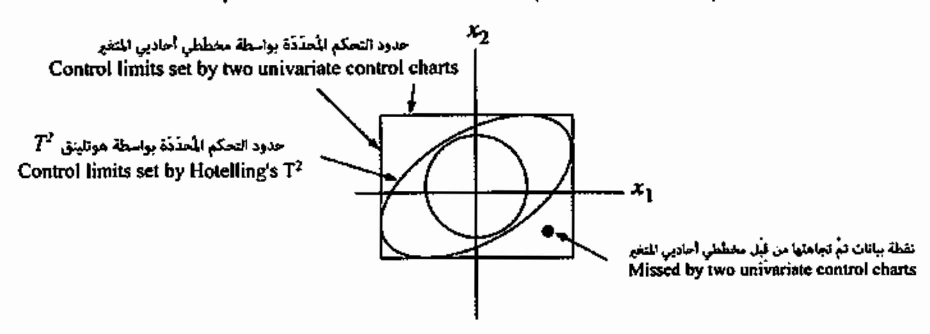
حيث إن S^{-1} هو معكوس المصفوفة S

تقيس إحصاءة هوتلينق T^2 المسافة الإحصائية لـ x_i من $\bar{x}$.

الشكل (١٧-١)

توضيح للمسافة الإحصائية المقاسة باستخدام إحصاءة هوتلينق T^2

وحدود التحكم لمخططات التحكم لهوتلينق T^2 ومخططات التحكم أحادية المتغير



لنفترض أن لدينا $\bar{x}=0$ عند نقطة الأصل من فضاء ثنائي الأبعاد لـ x_1 وx_2 في الشكل ١٧-١. في الشكل ١٧-١، تقع نقاط البيانات x_is بنفس المسافة الإحصائية من $\bar{x}$ داخل القطع الناقص (*ellipse*) أخذًا في الاعتبار التباين والتغاير لـ x_1 وx_2 في حين أن كل نقاط البيانات x_is بنفس المسافة الإقليدية تقع في الدائرة. كلما كانت قيمة إحصاءة هوتلينق T^2 أكبر لملحوظة بيانات مرصودة x_i كلما كانت المسافة الإحصائية x_i أكبر من $\bar{x}$.

يرصد مخطط التحكم لهوتلينق T^2 إحصاءة هوتلينق T^2 في المعادلة ١٧-٣. إذا كانت x_{i1} ،... x_{ip} تتبع توزيعًا طبيعيًا متعدد المتغيرات، فإن القيمة المحولة لإحصائية هوتلينق T^2:

$$\frac{n(n-p)}{p(n+1)(n-1)}T^2$$

تتبع توزيع F مع p وعدد $(n\text{-}p)$ من درجات الحرية (*degrees of freedom*). ولذلك، فإن قيمة F المصنفة والمجدولة على مستوى محدد من الأهمية، على سبيل المثال، $\alpha=0.05$ ، يمكن استخدامها باعتبارها نقطة إنذار أو حد الإشارة (*signal threshold*). إذا كانت القيمة المحولة لإحصاءة هوتلينق T^2 لملحوظة بيانات مرصودة x_i أكبر من حد الإشارة هذا، فإن مخطط تحكم هوتلينق T^2 يعطي إشارة إلى أن x_i نقطة شاذة. يمكن لمخطط التحكم لهوتلينق T^2 اكتشاف كل من تحولات المتوسط والارتباطات المقابلة (*Counter- relationships*). تُعدُ الارتباطات المقابلة انحرافات كبيرة عن تركيبة التغاير للمتغيرات.

ويوضح الشكل ١٧-١ حدود التحكم المحددة من قبل مخططي تحكم فرديين $\bar{x}$ لكل من x_1 وx_2 على التوالي، وحدود التحكم المحددة من قبل مخطط التحكم لهوتلينق T^2 على أساس المسافة الإحصائية. نظرًا لأن كل من مخططات التحكم الفردية $\bar{x}$ لـ x_1 وx_2 لا تحتوي بِنية التغاير لكل من x_1 وx_2 فإن ملحوظة البيانات المرصودة التي تنحرف عن بِنية التغاير لكل من x_1 وx_2 يتم تجاهلها في مخططات التحكم الفردية $\bar{x}$ ولكن يتم اكتشافها بواسطة مخطط التحكم لهوتلينق T^2 كما هو موضح في الشكل ١٧-١. لقد أشار ريان (*Ryan, 1989*) إلى أنَّ مخططات التحكم لهوتلينق T^2 هي أكثر حساسيةً للارتباطات المقابلة من تحولات المتوسط، على سبيل المثال، إذا كان هناك علاقة موجبة بين متغيرين ويحدث تحول المتوسط مع كلا المتغيرين ولكن في نفس الاتجاه للحفاظ على ارتباطهما، فقد لا تكتشف مخططات التحكم لهوتلينق T^2 تحول المتوسط (*Ryan, 1989*). تُعتبر مخططات التحكم لهوتلينق T^2 أيضًا حساسةً لفرضية التوزيع الطبيعية متعددة المتغيرات.

المثال ١٧-١

تحتوي مجموعة بيانات نظام التصنيع في الجدول ١٤-١، والمنسوخة في الجدول ١٧-١، على متغيري الخاصية، x_7 وx_8، في تسع حالات من أعطال الآلة الواحدة. يتم حساب المتجه المتوسط للعينة ومصفوفة التباين - التغاير في الفصل ١٤ ومعطاه فيما يلي. قم ببناء مخطط التحكم لهوتلينق T^2 لتحديد ما إذا كانت ملحوظة البيانات المرصودة الأولى $x=(x_7,x_8)=(1,0)$ عبارة عن ملاحظة شاذة.

$$\bar{x} = \begin{bmatrix} \bar{x}_7 \\ \bar{x}_8 \end{bmatrix} = \begin{bmatrix} \frac{5}{9} \\ \frac{4}{9} \end{bmatrix}$$

$$S = \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}$$

بالنسبة لملحوظة البيانات المرصودة الأولى $x=(x_7,x_8)=(1,0)$ نقوم بحساب قيمة إحصاءة هوتلينق T^2:

$$T^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) = \begin{bmatrix} 1-\frac{5}{9} & 0-\frac{4}{9} \end{bmatrix} \begin{bmatrix} 0.2469 & -0.1358 \\ -0.1358 & 0.2469 \end{bmatrix}^{-1} \begin{bmatrix} 1-\frac{5}{9} \\ 0-\frac{4}{9} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{4}{9} & -\frac{4}{9} \end{bmatrix} \begin{bmatrix} 5.8070 & 3.1939 \\ 3.1939 & 5.8070 \end{bmatrix} \begin{bmatrix} \frac{4}{9} \\ -\frac{4}{9} \end{bmatrix} = 0.1435.$$

وتكون قيمة T^2 المحوّلة:

$$\frac{n(n-p)}{p(n+1)(n-1)}T^2 = \frac{(9)(9-2)}{(2)(9+1)(9-1)}(0.1435) = 0.0502.$$

وتكون قيمة F المجدولة لـ $\alpha = 0.05$ مع ٢ و٧ من درجات الحرية تساوي ٤,٧٤، والتي يتم استخدامها كحد الإشارة. وحيث إن ٠,٠٥٠٢ >٤,٧٤، فإن مخطط التحكم لهوتلينق T^2 لا يعطي إشارة أن $x=(x_7,x_8)=(1,0)$ عبارة عن ملاحظة شاذة.

الجدول (١٧-١)

مجموعة البيانات لاكتشاف أعطال النظام مع اثنين من متغيرات الجودة x_7 وx_8

رقم الحالة - Instance (الآلة المعطلة – Faulty Machine)	x_7	x_8
1 (M1)	1	0
2(M2)	0	1
3(M3)	1	1
4(M4)	0	1
5(M5)	1	0
6(M6)	1	0
7(M7)	1	0
8(M8)	0	1
9(M9)	0	0

١٧-٢ مخططات تحكم المتوسط المتحرك الموزون الأسي متعددة المتغيرات (Multivariate EWMA Control Charts):

إن مخططات التحكم لهوتلينق T^2 عبارة عن نسخة متعددة المتغيرات لمخططات التحكم لـــ $\bar{x}$ أحادية المتغير في الفصل ١٦. وتُعدُّ مخططات تحكم المتوسط المتحرك الموزون الأسي $EWMA$ متعددة المتغيرات عبارة عن نسخة من مخطط تحكم المتوسط المتحرك الموزون الأسي $EWMA$ متعددة المتغيرات في الفصل ١٦. يقوم مخطط تحكم المتوسط

المتحرك الموزون الأسي *EWMA* المتعدد المتغيرات بمراقبة الإحصاءة التالية *(Ye, 2003, Chapter 4)*:

$$T^2 = z_i' S_z^{-1} z_i \;, \qquad (٤-١٧)$$

حيث إن:

$$z_i = \lambda x_i + (1-\lambda) z_{i-1} \;, \qquad (٥-١٧)$$

λ عبارة عن وزن في النطاق *[0,1]*،

$$z_0 = \mu \quad or \quad \bar{x} \qquad (٦-١٧)$$

$$S_z = \frac{\lambda}{\lambda-2}\left[1-(1-\lambda)^{2i}\right] S \qquad (٧-١٧)$$

وS هي مصفوفة تباين- تغاير العينة للمتغير x.

٣-١٧ مخططات تحكم مربع كاي (Chi-Square Control Charts):

حيث إن مخططات التحكم لهوتلينق T^2 ومخططات تحكم المتوسط المتحرك الموزون الأسي *EWMA* متعددة المتغيرات تتطلب حساب معكوس مصفوفة التباين- التغاير، فإن مخططات التحكم هذه ليست قابلة للقياس لعدد كبير من المتغيرات. إن وجود متغيرات مترابطة خطياً يخلق صعوبة في الحصول على معكوس مصفوفة التباين- التغاير. ولمعالجة هذه المشاكل، تمّ تطوير مخططات تحكم مربع كاي *(Ye et al., 2002b, 2006)*. يقوم

مخطط تحكم مربع كاي بمراقبة إحصاءة مربع كاي للملحوظة بيانات مرصودة $x_i = (x_{1i},..,$ $x_{ip})$ على النحو التالي:

$$\mathcal{X}^2 = \sum_{j=1}^{p} \frac{(x_{ij} - \bar{x}_j)^2}{\bar{x}_j}. \qquad (\text{٨-١٧})$$

على سبيل المثال، تضم مجموعة بيانات نظام التصنيع في الجدول ١٧-١ متغيري الخاصية، x_7، x_8 في تسع حالات من أعطال الآلة الأحادية. يتم حساب المتجه المتوسط للعينة في الفصل ١٤ ومعطى هنا:

$$\bar{x} = \begin{bmatrix} \bar{x}_7 \\ \bar{x}_8 \end{bmatrix} = \begin{bmatrix} \frac{5}{9} \\ \frac{4}{9} \end{bmatrix}$$

وتكون إحصائية مربع كاي للملحوظة البيانات المرصودة الأولى في الجدول ١٧-١، $(1, 0) =$ $x = (x_7, x_8)$:

$$\mathcal{X}^2 = \sum_{j=7}^{8} \frac{(x_{1j} - \bar{x}_j)^2}{\bar{x}_j} = \frac{(x_{17} - \bar{x}_7)^2}{\bar{x}_7} + \frac{(x_{18} - \bar{x}_8)^2}{\bar{x}_8} = \frac{\left(1 - \frac{5}{9}\right)^2}{\frac{5}{9}} + \frac{\left(0 - \frac{4}{9}\right)^2}{\frac{4}{9}} = 0.8.$$

إذا كانت المتغيرات التي عددها p مستقلةً وكانت قيمة p كبيرة، فإن إحصاءة مربع كاي تتبع توزيعًا طبيعيًا مبني على أساس نظرية النهاية المركزية. إذا كان لدينا عينة من ملحوظات البيانات المرصودة تحت السيطرة (*in- control*)، فإنه يمكن حساب متوسط العينة $\overline{\mathcal{X}^2}$ وتباين العينة $s_{\mathcal{X}^2}$ لاحصاءة مربع كاي واستخدامها لتحديد حدود التحكم:

$$UCL = \overline{\mathcal{X}^2} + L s_{\mathcal{X}^2} \qquad (\text{٩-١٧})$$

$$LCL = \overline{\mathcal{X}^2} - L s_{\mathcal{X}^2}. \qquad (\text{١٠-١٧})$$

إذا جعلنا $L=3$، يكون لدينا حدود تحكم ٣- سيغما. إذا كانت قيمة إحصاءة مربع كاي لملحوظة بيانات مرصودة معينة تقع خارج $[LCL, UCL]$، فإن مخطط تحكم مربع كاي يشير إلى حالة شاذة.

في العمل الذي أجراه يي وآخرون (*Ye et al., 2006*)، تتم مقارنة مخططات تحكم مربع كاي مع مخططات التحكم لهوتلينق T^2 في أدائهم لاكتشاف تحولات المتوسط والارتباطات المقابلة لأربعة أنواع من البيانات: (١) بيانات مع متغيرات مترابطة (*Correlated*) وموزعة بشكل طبيعي، (٢) بيانات مع متغيرات غير مترابطة وموزعة بشكل طبيعي، (٣) بيانات مع متغيرات مترابطة ذاتيا (مع نفسها) وموزعة بشكل طبيعي، و(٤) متغيرات موزعة بشكل غير طبيعي وبدون ارتباط مع متغيرات أخرى أو ارتباط مع نفسها. تُظهِر نتائج الاختبارات أن أداء مخططات تحكم مربع كاي كان هو الأفضل أو بنفس جودة أداء مخططات التحكم لهوتلينق T^2 للبيانات من الأنواع ٢ و٣ و٤. كان أداء مخططات التحكم لهوتلينق T^2 أفضل من مخططات تحكم مربع كاي للبيانات من النوع ١ فقط. لكن، بالنسبة للبيانات من النوع ١، يمكننا استخدام تقنيات مثل تحليل المكون الرئيسي (*principal component analysis*) في الفصل ١٤ للحصول على المكونات الرئيسية. ثم يمكن استخدام مخطط تحكم مربع كاي لمراقبة المكونات الرئيسية التي هي عبارة عن متغيرات مستقلة.

٤-١٧ التطبيقات (Applications):

يمكن إيجاد تطبيقات لمخططات التحكم لهوتلينق T^2 ومخططات تحكم مربع كاي لاكتشاف الهجوم الإلكتروني/عبر الإنترنت لرصد بيانات الحاسب والشبكات واكتشاف الهجمات الإلكترونية كحالات شاذة في العمل الذي أجراه يي وزملاؤه (*Emran and Ye, 2002; Ye, 2003, Chapter 4; Ye, 2008; Ye and Chen, 2001; Ye et al., 2001, 2003, 2004, 2006*). وهناك أيضًا تطبيقات لمخططات تحكم متعدد المتغيرات في التصنيع (*Ye, 2003, Chapter 4*) وغيرها من المجالات.

التمارين (Exercises):

١-١٧ قم باستخدام مجموعة البيانات x_4 x_5 وx_6 في الجدول ٨-١ لتقدير المعلَمات لمخطط التحكم لهوتلينق T^2 ثم قم ببناء مخطط التحكم لهوتلينق T^2 مع $\alpha = 0.05$ لمجموعة البيانات x_4 x_5 وx_6 في الجدول ٤-٦ لرصد البيانات واكتشاف أي حالات شاذة.

٢-١٧ قم باستخدام مجموعة البيانات x_4 x_5 وx_6 في الجدول ٨-١ لتقدير المعلَمات لمخطط تحكم مربع كاي ثم قم ببناء مخطط تحكم مربع كاي مع $L=3$ لمجموعة البيانات x_4 x_5 وx_6 في الجدول ٤-٦ لرصد البيانات واكتشاف أي حالات شاذة.

٣-١٧ كرر المثال ١٧-١ لملحوظات البيانات المرصودة الثانية.

الجزء السادس
خوارزميات استكشاف الأنماط الزمنية والتسلسلية
Algorithms for Mining Sequential and Temporal Patterns

١٨- تحليل الارتباط الذاتي والسلاسل الزمنية
Autocorrelation and Time Series Analysis

تتكون بيانات سلاسل الزمن (*Time Series data*) من مشاهدات (أو ملحوظات) لبيانات يتم رصدها على مدى زمني معين. فإذا أصبحت ملحوظات البيانات المرصودة مترابطةً مع بعضها على مدى زمني فإنه يمكن القول إن بيانات السلاسل الزمنية مترابطة ذاتيا (*autocorrelated*). تمّ تقديم تحليل سلاسل الزمن بواسطة بوكس وجنكينز سنة ١٩٧٦ (*Box and Jenkins, 1976*) لنمذجة وتحليل بيانات سلاسل الزمن ذات الارتباط الذاتي. وقد تمّ تطبيق تحليل سلاسل الزمن على بيانات حقيقية في العديد من المجالات، بما في ذلك أسعار الأسهم (على سبيل المثال، مؤشر *S & P 500*)، وأجرة تذاكر الطيران، وحجم القوى العاملة، وبيانات البطالة، وأسعار الغاز الطبيعي (*Yaffee and McGee, 2000*). يوجد بيانات سلاسل زمنية ساكنة (*stationary*) وغير ساكنة (*nonstationary*) والتي تتطلب إجراءات مختلفة للاستدلال الإحصائي. في هذا الفصل، يتم تعريف الارتباط الذاتي (*autocorrelation*). ويتم توضيح عدة أنواع من السلاسل الزمنية الساكنة وغير الساكنة. ويتم توصيف نماذج المتوسط المتحرك ذاتي الانحدار (*Autoregressive and Moving Average - ARMA*) الخاصة ببيانات السلاسل الساكنة. ويتم استعراض عملية تحويل بيانات السلاسل غير الساكنة إلى بيانات سلاسل ساكنة، جنبًا إلى جنب مع نماذج المتوسط المتحرك، المتكاملة، وذاتية الانحدار (*Autoregressive, Integrated, Moving Average - ARIMA*). وترد قائمة من حزم البرمجيات التي تدعم تحليل السلاسل الزمنية. يتم تقديم بعض التطبيقات الخاصة بتحليل السلاسل الزمنية مع المراجع الخاصة بها.

١٨-١ الارتباط الذاتي (Autocorrelation):

تقدم المعادلة ١٤-٧ في الفصل ١٤ معامل الارتباط (*coefficient correlation*) لمتغيرين x_i وx_j:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}},$$

حيث تعطي المعادلتان ١٤-٤ و١٤-٦،

$$\sigma_i^2 = \sum_{\substack{all\ valuese \\ of\ x_i}} (x_i - u_i)^2 p_i(x_i)$$

$$\sigma_{ij} = \sum_{\substack{all\ valuese \\ of\ x_i}} \sum_{\substack{all\ valuese \\ of\ x_j}} (x_i - \mu_i)(x_j - \mu_j) p_i(x_i, x_j).$$

إذا كان لدينا متغير x وعينة من بيانات السلاسل الزمنية الخاصة بالمتغير ولتكن x_t بحيث $t = 1, \ldots, n$، فإننا نحصل على معامل دالة الارتباط الذاتي بفارق زمني k (*the k lag-k autocorrelation function [AFC] coefficient*) عن طريق استبدال المتغيرين x_i وx_j في المعادلات المذكورة أعلاه بالمتغيرين x_t وx_{t-k} وهما ملاحظتا بيانات مرصودتان بفارق زمني k:

$$\text{ACF}(k) = \rho_k = \frac{\sum_{t=k+1}^{n} (x_t - \bar{x})(x_{t-k} - \bar{x})/(n-k)}{\sum_{t=1}^{n} (x_t - \bar{x})^2/n}, \qquad (١-١٨)$$

حيث $\bar{x}$ هو متوسط العينة. إذا كانت بيانات السلاسل الزمنية مستقلةً إحصائيًا عند فارق الزمن k (*lag-k*)، يكون ρ_k بقيمة صفر. إذا تغير x_t وx_{t-k} من المتوسط $\bar{x}$ بنفس الاتجاه (على سبيل المثال، كلاهما يزيدان من $\bar{x}$)، تكون ρ_k موجبة. إذا تغيرت x_t وx_{t-k} من المتوسط $\bar{x}$ باتجاه معاكس (على سبيل المثال، تزيد واحدة وتنقص الأخرى من المتوسط $\bar{x}$)، تكون ρ_k سالبة.

يقوم معامل دالة الارتباط الذاتي الجزئي بفارق زمني k (*Partial Autocorrelation Function - PACF*) بقياس الارتباط الذاتي للفارق الزمني k، والذي لا يؤخَذ به في الاعتبار من قبل الارتباطات الذاتية للفوارق الزمنية من 1 إلى $k-1$. وتوضح المعادلة التالية دالة الارتباط الذاتي الجزئي (*PACF*) للفارق الزمني 1، (*lag-1*)، وللفارق الزمني 2، (*lag-2*)، (*Yaffee and McGee, 2000*):

$$\text{PACF}(1) = \rho_1 \quad (٢-١٨)$$

$$\text{PACF}(2) = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}. \quad (٣-١٨)$$

٢-١٨ السكون واللاسكون (Stationarity and Nonstationarity):

عادةً ما يشير السكون إلى سكون ضعيف يتطلب أن لا يتغير المتوسط (*mean*) والتباين (*variance*) الخاص ببيانات السلاسل الزمنية مع مرور الوقت. تكون السلسلة الزمنية ساكنةً بشكل دقيق إذا كان التغاير الذاتي $\sigma_{t,t-k}$، لا يتغير بمرور الوقت t، ولكن يعتمد فقط على العدد k، الذي يمثل الفارق الزمني، بالإضافة إلى المتوسط الثابت والتباين الثابت. على سبيل المثال، إنَّ سلسلة قوسشيان الزمنية (*Gaussian time series*) التي لها توزيع طبيعي متعدد المتغيرات هي عبارة عن سلسلة ساكنة بشكل دقيق وصارم لأن المتوسط، والتباين، والتغاير الذاتي للسلسلة (*autocovariance*) لا تتغير مع مرور الوقت. وتُستخدَم نماذج المتوسط المتحرك ذاتي الانحدار (*ARMA*) لنمذجة السلاسل الزمنية الساكنة.

قد يكون السبب في اللاسكون (*Nonstationarity*) هو:

- الحالات المتطرفة (*outliers*) (انظر الوصف في الفصل ١٦).
- السير العشوائي (*random walk*) والذي فيه تنحرف كل ملحوظة من ملحوظات البيانات المرصودة بشكل عشوائي من ملحوظة البيانات المرصودة السابقة دون الرجوع إلى المتوسط.
- الاتجاه المحدد (*deterministic trend*) (على سبيل المثال، اتجاه خطي - *linear trend*- له قيم تتغير مع مرور الوقت بمعدل ثابت ومستمر).
- التباين المتغير.
- تكرار نمط بيانات معين بشكل دوري (دورة نمط بيانات)، بما في ذلك الدورات الموسمية بشكل سنوي.
- أسباب أخرى تجعل المتوسط أو التباين للسلسلة الزمنية تتغير بمرور الزمن.

يجب أن يتم تحويل السلسلة غير الساكنة إلى سلسلة ساكنة من أجل بناء نموذج المتوسط المتحرك ذاتي الانحدار (*Autoregressive and Moving Average - ARMA*).

٣-١٨ نماذج المتوسط المتحرك ذاتي الانحدار الخاصة ببيانات السلاسل الساكنة: (ARMA Models of Stationary Data)

يتم تطبيق نماذج المتوسط المتحرك ذاتي الانحدار (*Autoregressive and Moving Average - ARMA*) على بيانات السلاسل الزمنية ذات السكون الضعيف. يقوم نموذج الانحدار الذاتي (*Auto Regressive-AR*) ذو الدرجة *p*، *AR(p)* بوصف السلسلة الزمنية التي تكون فيها ملحوظة البيانات المرصودة الحالية لمتغير xهي دالة لعدد p من ملحوظاتها المرصودة السابقة، وخطأ عشوائي:

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + e_t. \qquad (٤\text{-}١٨)$$

على سبيل المثال، يتم نمذجة بيانات السلاسل الزمنية لمدى استحسان الأداء الوظيفي للرئيس استنادًا إلى استطلاع غالوب كنموذج انحدار ذاتي من الدرجة (*P=1*) وتُكتَب *AR(1)* (*Yaffee and McGee, 2000*):

$$x_t = \phi_1 x_{t-1} + e_t. \qquad (٥\text{-}١٨)$$

يوضح الجدول ١٨-١ سلسلة زمنية لنموذج انحدار ذاتي *AR(1)* حيث $\phi_1 = 0.09$، و$x_0 = 3$ وخطأ عشوائي e_t ذو متوسط يساوي صفراً، وانحراف معياري يساوي واحداً.

يوضح الشكل ١٨-١ رسمًا بيانياً لسلسلة زمنية بنموذج انحدار ذاتي *AR(1)*. كما نرى في الشكل ١٨-١، فإن تأثير قيمة x الأولية، $x_0 = 3$ ينعدم بسرعة. يقوم نموذج المتوسط المتحرك (*Moving Average- MA*) من الدرجة *q*، *MA(q)*، بوصف سلسلة زمنية والتي فيها ملحوظة البيانات المرصودة الحالية لمتغير معين عبارة عن تأثير خطأ عشوائي في الوقت الحالي والأخطاء العشوائية لعدد q من نقاط زمنية سابقة:

$$x_t = e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}. \qquad (٦-١٨)$$

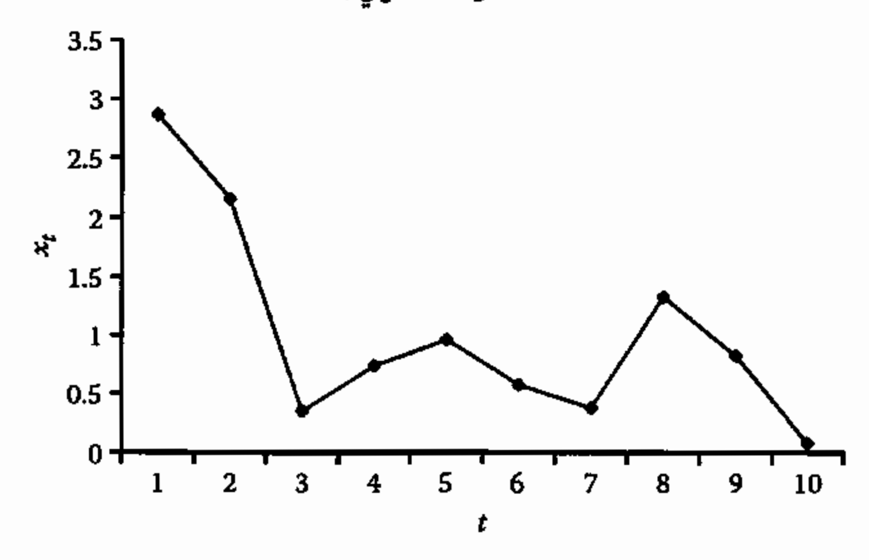
الجدول (١٨-١)

سلسلة زمنية لنموذج الانحدار الذاتي $AR(1)$ حيث $\phi_1 = 0.09$، $x_0 = 3$ وخطأ عشوائي e_t

x_t	e_t	t
2.866	0.166	1
2.157	−0.422	2
0.353	−1.589	3
0.741	0.424	4
0.962	0.295	5
0.579	−0.287	6
0.381	−0.140	7
1.328	0.985	8
0.825	−0.370	9
0.078	−0.665	10

الشكل ١٨-١

بيانات سلسلة زمنية يتم توليدها باستخدام نموذج الانحدار الذاتي $AR(1)$ حيث $\phi_1 = 0.09$، و $x_0 = 3$، وخطأ عشوائي e_t



على سبيل المثال، يتم نمذجة بيانات السلسلة الزمنية الخاصة بتتبع المصابين بمرض وبائي كنسبة من مجموعة سكانية مصابة بمرض بشكل عام (مثل، الإيدز) كنموذج متوسط متحرك، *MV(1)* (*Yaffee and McGee, 2000*)

$$x_t = e_t - \theta_1 e_{t-1}. \qquad (٧\text{-}١٨)$$

يقدم الجدول ١٨-٢ سلسلة زمنية لنموذج المتوسط المتحرك *MV(1)* حيث $\theta_1 = 0.9$، وخطأ عشوائيا e_t بمتوسط يساوى صفرا، وانحرافًا معياريًا يساوي واحدًا. يوضح الشكل ١٨-٢ رسمًا بيانيا لسلسلة زمنية بنموذج المتوسط المتحرك *MV(1)*. كما نرى في الشكل ١٨-٢، فإن قيمة $(-0.9e_{t-1})$ في المعادلة ١٨-٧ تميل إلى أخذ x_t إلي الاتجاه المعاكس من x_{t-1} مما يجعل قيم x_t تتأرجح.

يقوم نموذج المتوسط المتحرك ذاتي الانحدار *ARMA*، ونموذج المتوسط المتحرك ذاتي الانحدار *ARMA*(p,q)، بوصف سلسلة زمنية بخصائص المتوسط المتحرك، وذاتي الانحدار:

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + e_t - \theta_1 x_{t-1} - \cdots - \theta_q x_{t-q}. \qquad (٨\text{-}١٨)$$

يرمز الرمز *ARMA* $(p, 0)$ إلى نموذج الانحدار الذاتي *AR(p)* والرمز *ARMA* $(0,q)$ إلى نموذج المتوسط المتحرك *MA(q)*. بشكل عام، يكون لسلسلة الزمن السـلـسـة (*smooth time series*) معاملات (*coefficients*) انحدار ذاتي *AR* عالية، ومعاملات متوسط متحرك *MA* منخفضة، ويكون للسلسلة الزمنية المتأثرة بالأخطاء العشوائية معاملات متوسط متحرك *MA* عالية، ومعاملات انحدار ذاتي *AR* منخفضة.

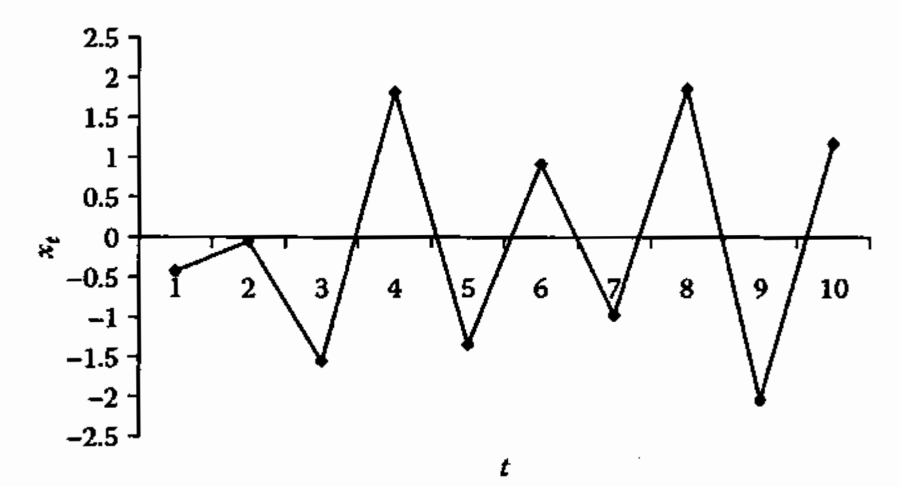
الجدول (١٨-٢)

سلسلة زمنية لنموذج $MA(1)$ مع $\theta_1 = 0.9$ وخطأ عشوائي e_t

x_t	e_t	t
	0.649	0
−0.418	0.166	1
−0.046	−0.422	2
−1.548	−1.589	3
1.817	0.424	4
−1.340	0.295	5
0.919	−0.287	6
−0.967	−0.140	7
1.856	0.985	8
−2.040	−0.370	9
1.171	−0.665	10

الشكل (١٨-٢)

بيانات سلسلة زمنية تمّ توليدها باستخدام نموذج $MA(1)$ مع $\theta_1 = 0.9$ وخطأ عشوائي e_t



١٨-٤ خصائص دالة الارتباط الذاتي ودالة الارتباط الذاتي الجزئي لنماذج المتوسط المتحرك ذاتي الانحدار
(ACF and PACF Characteristics of ARMA Models):

تقوم دالة الارتباط الذاتي (*Autocorrelation Function - ACF*)، ودالة الارتباط الذاتي الجزئي (*Partial Autocorrelation Function - PACF*) التي تمّ وصفها في الجزء ١٨-١ بتوفير الأدوات التحليلية لكشف وتحديد درجة الانحدار الذاتي (*AR*)، أو درجة المتوسط المتحرك (*MA*) في نموذج المتوسط المتحرك ذاتي الانحدار (*ARMA*) لسلسلة زمنية. فيما يلي، يتم توضيح خصائص كل من *ACF*، *PACF* لبيانات السلاسل الزمنية التي تمّ توليدها بواسطة نماذج الانحدار الذاتي *AR*، والمتوسط المتحرك *MA*، والمتوسط المتحرك ذاتي الانحدار *ARMA*.

بالنسبة لسلسلة زمنية بانحدار ذاتي من الدرجة ١، *AR(1)*:

$$x_t = \phi_1 x_{t-1} + e_t \, ,$$

تكون دالة الارتباط الذاتي (*Yaffee and McGee, 2000*) *ACF(k)*:

$$\mathrm{ACF}(k) = \phi_1^k. \qquad (٩\text{-}١٨)$$

إذا كان $\phi_1 < 1$، فإنّ *AR(1)* يكون ساكنًا ويتراجع أسّي في القيمة المطلقة لـــــ *ACF* مع مرور الوقت لأن *ACF(k)* يتناقص بمقدار k ويتلاشى في النهاية. إذا كان $\phi_1 > 0$، فإن *ACF(k)* يكون موجباً. إذا كان $\phi_1 < 0$، فإن *ACF(k)* تتأرجح بحيث تكون سالبة بالنسبة لـ $k = 1$، وموجبة بالنسبة لـ $k = 2$، وسالبة بالنسبة لـ $k = 3$، وموجبة بالنسبة لـ $k = 4$، وهلّم جرا. إذا كان $\phi_1 \geq 1$، فإن *AR(1)* يكون غير ساكن. بالنسبة لسلسلة زمنية ساكنة بانحدار ذاتي من الدرجة ٢، *AR(2)*:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + e_t \, ,$$

فإن *ACF(k)* تكون موجبة بتراجع أسّي. في القيمة المطلقة لـ *ACF*مع مرور الوقت، إذا كان $\phi_1 > 0$ و $\phi_2 > 0$، وتتأرجح قيمة *ACF(k)* بتراجع أسّي في القيمة المطلقة لـ *ACF*مع مرور الوقت إذا كان $\phi_1 < 0$ و $\phi_2 > 0$.

تنتهي دالة الارتباط الذاتي الجزئي *PACF(k)* لسلسلة انحدار ذاتي *AR(p)*بإكمال الفارق الزمني *p*، وتصبح صفرا بعد فارق زمني *p*. بالنسبة لـ *AR(1)*، فإن *PACF (1)* تكون موجبة إذا كان $\phi_1 > 0$ أو سالبة إذا كان $\phi_1 < 0$، وتكون *PACF(k)* لـ $k \geq 2$ مساوية للصفر. وبالنسبة لـ *AR(2)*، فإن*PACF(1)*، و*PACF(2)* تكون موجبة إذا كان $\phi_2 > 0$و $\phi_1 > 0$، وتكون *PACF(1)* سالبة، و*PACF(2)* موجبة إذا كان $\phi_1 < 0$ و $\phi_2 > 0$، وتكون *PACF(k)* لـ $k \geq 3$ مساويةً للصفر. وبالتالي، فإنَّ *PACF* تحدد درجة سلسلة الزمن ذاتية الانحدار.

بالنسبة للسلسلة الزمنية ذات *MA(1)*،

$$x_t = e_t - \theta_1 e_{t-1},$$

فإن *ACF*(١) لا تكون صفرًا كما يلي (*Yaffee and McGee, 2000*):

$$\mathrm{ACF}(1) = \frac{-\theta_1}{1+\theta_1^2}, \qquad (١٠-١٨)$$

وتكون *ACF(k)* صفراً بالنسبة لـ $k > 1$. بالمثل للسلسلة الزمنية ذات *MA (2)*، فإن*ACF(1)*، و*ACF(2)* تكون سالبة، و*ACF(q)* تساوي صفراً لـ $q > 2$. وبالنسبة لـ *MA(q)*، يكون لدينا (*Yaffee and McGee, 200*):

$$\begin{aligned} \mathrm{ACF}(k) &\neq 0 \quad \text{if } k \leq q \\ \mathrm{ACF}(k) &= 0 \quad \text{if } k > q \end{aligned}$$

خلافاً لسلسلة الزمن ذاتية الانحدار التي تنخفض دالة الارتباط الذاتي ACF الخاصة بها بشكل أسّي بمرور الوقت، فإن السلسلة الزمنية للمتوسط المتحرك يكون لها ذاكرة محدودة لأن الارتباط الذاتي لـ $MA(q)$ ينتهي بإكمال الفارق الزمني q. وبالتالي، تقوم دالة الارتباط الذاتي ACF بتحديد درجة السلسلة الزمنية للمتوسط المتحرك. والسلسلة الزمنية للمتوسط المتحرك يكون لها دالة $PACF$ والتي ينخفض حجمها بشكل أسّي مع مرور الوقت. بالنسبة لـ $MA(1)$، فإن $PACF(k)$ تكون سالبة إذا كان $\theta_1 > 0$، وتتأرجح $PACF(k)$ بين القيم الموجبة والسالبة وبتراجع أسّي في حجم $PACF(k)$ مع مرور الوقت. بالنسبة $MA(2)$، فإن $PACF(k)$ تكون سالبة وبتراجع أسّي في حجم $PACF$ مع مرور الوقت إذا كان $\theta_1 > 0$ و $\theta_2 > 0$، وتتأرجح قيمة $ACF(k)$ بتراجع أسّي في القيمة المطلقة لـ ACF بمرور الوقت إذا كان $\theta_1 < 0$ و $\theta_2 < 0$.

يتم الجمع بين الخصائص المذكورة آنفًا والخاصة بالسلاسل الزمنية ذات المتوسط المتحرك وذاتية الانحدار في سلسلة زمنية مختلطة بنماذج $ARMA(p, q)$ حيث $p > 0$ و $q > 0$، فعلى سبيل المثال، بالنسبة لـ $ARMA(1,1)$ مع $\phi_1 > 0$ و $\theta_1 < 0$، تنخفض دالة الارتباط الذاتي ACF بشكل أسّي بمرور الوقت، وتتأرجح دالة الارتباط الذاتي الجزئي $PACF$ بتراجع أسّي بمرور الوقت.

يمكن تقدير المعلَمات في نموذج المتوسط المتحرك ذاتي الانحدار $ARMA$ من عينة بيانات السلسلة الزمنية باستخدام طريقة المربعات الصغرى غير المشروطة (*unconditional least-squares method*)، طريقة المربعات الصغرى المشروطة، أو طريقة الإمكان الأكبر (*Yaffee and McGee, 2000*)، والتي يتم دعمها في البرامج الإحصائية، مثل: *SAS* (*www.sas.com*) و*SPSS* (*www.ibm.com/software/analytics/spss/*).

١٨-٥ تحويل بيانات السلسلة غير الساكنة ونماذج المتوسط المتحرك المتكامل ذاتي الانحدار (Transformations of Nonstationary Series Data and ARIMA Models):

بالنسبة للسلسلة غير الساكنة الناجمة عن القيم المتطرفة والشاذة، والسير العشوائي، والاتجاه المحدد، والتباين المتغير، والتكرار الدوري والموسمي، والتي تمّ وصفها في الجزء ١٨-٢، يتم فيما يلي وصف الطرق الخاصة بتحويل تلك السلسلة غير الساكنة إلى سلسلة ساكنة.

عندما يتم الكشف عن القيم المتطرفة والشاذة في سلسلة زمنية، فإنه من الممكن أن يتم إزالتها واستبدالها، وذلك باستخدام متوسط هذه السلسلة. وتنحرف كل ملحوظة بيانات عشوائيًا في السير العشوائي من ملحوظة البيانات السابقة دون الرجوع إلى المتوسط. السائقون المخمورون ومعدلات المواليد عبارة عن أسئلة تمثل سلوك السير العشوائي (*Yaffee and McGee, 2000*). يتم تطبيق يتم تطبيق عملية الطرح على سلسلة السير العشوائي على النحو التالي:

$$e_t = x_t - x_{t-1} \qquad (١١-١٨)$$

للحصول على سلسلة ساكنة من المتبقي e_t، والتي يتم بعد ذلك نمذجتها كنموذج متوسط متحرك ذاتي الانحدار *ARMA* يمكن إزالة اتجاه محدد معين مثل الاتجاه الخطي التالي:

$$x_t = a + bt + e_t\ , \qquad (١٢-١٨)$$

عن طريق إعادة التوجيه (*de-trending*). يتضمن إعادة التوجيه أولًا القيام ببناء نموذج انحدار للتعرف على الاتجاه (على سبيل المثال، نموذج خطي لاتجاه خطي، أو نموذج متعدد الحدود للاتجاه ذو الدرجة الأعلى) ومن ثمّ الحصول على السلسلة الساكنة من البواقي e_t من خلال إجراء عملية الطرح بين القيمة المرصودة والقيمة المتوقعة من نموذج الانحدار. بالنسبة للتباين المتغير (*changing variance*) الذي له تباين سلسلة زمنية ممتدة، أو منكمشة، أو متذبذبة، مع مرور الوقت، فإنه من الممكن إجراء التحويل باستخدام اللوغاريتم الطبيعي (*natural log*) أو التحويل باستخدام الرفع للقوة (على سبيل المثال، التربيع والجذر التربيعي) لتحقيق الاستقرار في التباين (*Yaffee and McGee, 2000*). تنتمي التحويلات اللوغاريتمية الطبيعية، أو تحويلات القوة إلى عائلة تحويلات بوكس-كوكس (*Box-Cox*)، التي تُعرف بأنها (*Yaffee and McGee, 2000*):

$$y_t = \frac{(x_t + c)^{\lambda} - 1}{\lambda} \quad \text{if } 0 < \lambda \leq 1 \qquad (١٣-١٨)$$
$$y_t = \ln x_t + c \quad \text{if } \lambda = 1$$

حيث:

x_t السلسلة الزمنية الأصلية

y_t السلسلة الزمنية المتحولة

c ثابت

λ معلَمة شكل (*shape parameter*)

بالنسبة للسلسلة الزمنية المكونة من تكرارات دورية (*cycles*)، والتي يكون بعضها موسمياً بدورة سنوية، يمكن إجراء عملية طرح دورية أو موسمية على النحو التالي:

$$e_t = x_t - x_{t-d} \quad (١٨-١٤)$$

حيث إن d هو عدد مرات الفوارق الزمنية الممتدة عبر الدورة. يمكن إضافة عملية الطرح العادية وعملية الطرح الدورية/ الموسمية إلى نموذج *ARMA* ليصبح نموذج المتوسط المتحرك، المتكامل، وذاتي الانحدار (*Autoregressive, Integrated, Moving Average - ARIMA*) حيث تشير I إلى الكلمة متكامل (*Integrated*):

$$x_t - x_{t-d} = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + e_t - \theta_1 x_{t-1} - \cdots - \theta_q x_{t-q}. \quad (١٨-١٥)$$

١٨-٦ البرمجيات والتطبيقات (Software and Applications):

يتم دعم تحليل السلاسل الزمنية بمجموعة من الحزم البرمجية مثل SAS (www.sas.com)، وSPSS (www.ibm.com/software/analytics/spss/) و*MATLAB (www.mathworks.com)*. في العمل الذي قامت به يي وزملاؤها (*Ye, 2008, Chapter 10 and 17*)، يتم تطبيق تحليل السلاسل الزمنية لكشف وتحديد خصائص الارتباط الذاتي للاستخدام العادي وأنشطة الهجوم عبر الإنترنت باستخدام بيانات الحاسوب والشبكات. يتم بناء نماذج السلاسل الزمنية على أساس هذه الخصائص ويتم استخدامها في مخططات تحكم الدرجة التراكمية (*cuscore*) كما هو موضح في الفصل ١٦ للكشف عن وجود هجمات إلكترونية. يمكن العثور على التطبيقات الخاصة بتحليل السلاسل الزمنية بغرض التنبؤ في يافي وماغي (*Yaffee and McGee, 2000*).

التمارين (Exercises):

١٨-١ قم ببناء بيانات سلاسل زمنية باستخدام نموذج $ARMA(1,1)$.

١٨-٢ بالنسبة لبيانات السلاسل الزمنية في الجدول ١٨-١، قم بحساب $ACF(1)$، $ACF(2)$، $ACF(3)$، $PACF(1)$، و$PACF(2)$.

١٨-٣ بالنسبة لبيانات السلاسل الزمنية في الجدول ١٨-٢، قم بحساب $ACF(1)$، $ACF(2)$، $ACF(3)$، $PACF(1)$، و $PACF(2)$.

١٩- نماذج سلسلة ماركوف ونماذج ماركوف المخفية
Markov chain Models and Hidden Markov Models

يتم استخدام نماذج سلسلة ماركوف ونماذج ماركوف المخفية على نطاق واسع لبناء النماذج، ولعمل الاستدلالات والاستنتاجات الخاصة بأنماط البيانات المتعاقبة. في هذا الفصل، يتم وصف نماذج سلسلة ماركوف ونماذج ماركوف المخفية. وترد قائمة من حزم البرمجيات لاستكشاف البيانات التي تدعم التعلُم والاستدلال من نماذج سلسلة ماركوف ونماذج ماركوف المخفية. ويتم إعطاء بعض التطبيقات من نماذج سلسلة ماركوف ونماذج ماركوف المخفية مع المراجع.

١-١٩ نماذج سلسلة ماركوف: (Markov Chain Models)

يصف نموذجُ سلسلة ماركوف العمليةَ العشوائية أو التصادفية (*stochastic process*) بأوقات منفصلة (*discrete-time*) ومن الدرجة الأولى (*frist-order*) لنظام له خاصية ماركوف والمتعلقة باحتمال أن حالة النظام (*system state*) في الوقت n لا تعتمد على حالات النظام السابقة، المؤدية إلى حالة النظام في وقت $n - 1$، ولكن فقط على حالة النظام عند $n - 1$:

$$P(s_n|s_{n-1},\dots,s_1) = P(s_n|s_{n-1}) \quad \text{for all } n, \qquad (١-١٩)$$

حيث إن S_n هي حالة النظام في الوقت n. ويوجد خاصية إضافية لنموذج سلسلة ماركوف الساكنة (*stationary*) وهي أن احتمال انتقال الحالة من الوقت $n - 1$ إلى n هو مستقل عن الوقت n:

$$P(s_n = j|s_{n-1} = i) = P(j|i), \qquad (٢-١٩)$$

حيث إن $p(j|i)$ هو احتمال أن يكون النظام في الحالة j في وقت معين علمًا بأن النظام كان في الحالة i في الوقت السابق. وللتبسيط فإننا نطلق على نموذج ماركوف الساكن بنموذج ماركوف في هذا الكتاب.

إذا كان للنظام عدد محدود من الحالات، $1, \ldots, S$، فإنه يتم تعريف نموذج سلسلة ماركوف من خلال احتمالات انتقال أو تحول الحالة، $P(j|i)$، حيث إن: $i = 1, \ldots, S$، و $j = 1, \ldots, S$:

$$\sum_{j=1}^{S} P(j|i) = 1, \quad (١٩-٣)$$

واحتمالات الحالة الأولية، $P(i)$، حيث إن: $i = 1, \ldots, S$:

$$\sum_{j=1}^{S} P(i) = 1, \quad (١٩-٤)$$

حيث إن $P(i)$ هو احتمال أن يكون النظام في الحالة i في الوقت 1. يتم حساب الاحتمال المشترك لتسلسل معطى لحالات النظام $S_{n-K+1}, \ldots, S_n$ في إطار زمني طوله K بما في ذلك الأوقات المنفصلة $n-(K-1), \ldots, n$ على النحو التالي:

$$P(s_{n-K+1}, \ldots, s_n) = P(s_{n-K+1}) \prod_{k=K-1}^{1} P(s_{n-k+1}|s_{n-k}) \quad (١٩-٥)$$

يمكن تعلُّم واستخلاص احتمالات انتقال الحالة، واحتمالات الحالة الأولية من مجموعة البيانات التدريبية أو الاستكشافية التي تحتوي على واحد أو أكثر من تعاقب الحالات على النحو التالي:

$$P(j|i) = \frac{N_{ji}}{N_{.i}} \quad (٦\text{-}١٩)$$

$$P(i) = \frac{N_i}{N}, \quad (٧\text{-}١٩)$$

حيث إن:

N_{ji}	هو التكرار الذي يظهر فيه الانتقال من الحالة i إلى الحالة j في البيانات التدريبية
$N_{.i}$	هو من التكرار الذي يظهر فيه الانتقال من الحالة i إلى أي من الحالات، 1، ...، S في البيانات التدريبية
N_i	هو تكرار ظهور الحالة i في البيانات التدريبية
N	هو العدد الإجمالي للحالات في البيانات التدريبية

يمكن استخدام نماذج سلسلة ماركوف لمعرفة وتصنيف أنماط البيانات والمتعاقبة. لكل فئة من الفئات المستهدفة (*target class*)، يمكن استخدام البيانات المتعاقبة بالفئة المستهدفة لبناء نموذج سلسلة ماركوف عن طريق تعلم المصفوفة الاحتمالية لانتقال الحالة (*state transition probability matrix*)، والتوزيع الاحتمالي المبدئي من البيانات التدريبية وفقًا للمعادلات ١٩-٦ و١٩-٧. وهو ما يعني، أننا نحصل على نموذج سلسلة ماركوف لكل فئة من الفئات المستهدفة. إذا كان لدينا الفئات المستهدفة، 1، ...، c، فإننا نقوم ببناء نماذج سلسلة ماركوف، M_1، ...، M_c، لهذه الفئات المستهدفة. إذا كان لدينا سلسلة اختبارية، يتم حساب الاحتمال المشترك لهذه السلسلة باستخدام المعادلة ١٩-٥ تحت كل نموذج من نماذج سلسلة ماركوف. ويتم تصنيف السلسلة الاختيارية إلى الفئة المستهدفة لنموذج سلسلة ماركوف التي تعطي أعلى قيمة للاحتمال المشترك الخاص بالسلسلة الاختبارية.

في تطبيقات نماذج سلسلة ماركوف بغرض الكشف عن الهجمات الإلكترونية (*Ye et al., 2002c, 2004*)، يتم جمع بيانات التدقيق الحاسوبية، لحالات الاستخدام العادي، وحالات الهجمات الإلكترونية المتنوعة، على أجهزة الحاسوب. هناك ما مجموعه ٢٨٤ نوعًا

من أنواع أحداث التدقيق (*audit event*) في بيانات التدقيق. يتم اعتبار كل حدث من أحداث التدقيق واحدًا من ٢٨٤ حالة نظام. ويتم اعتبار كل حالة من الحالات (الاستخدام العادي والهجمات المختلفة) كفئة من الفئات المستهدفة (*target class*). يتم تعلم نموذج سلسلة ماركوف لفئة مستهدفة من البيانات التدريبية حسب حالة الفئة المستهدفة. لكل سلسلة اختبارية من أحداث التدقيق في إطار رصد معين، يتم حساب الاحتمال المشترك للسلسلة الاختبارية في إطار كل نموذج من نماذج سلسلة ماركوف. ويتم تصنيف السلسلة الاختبارية إلى أحد الحالات: (استخدام عادي، أو أحد أنواع الهجمات الإلكترونية) لتحديد ما إذا كان الهجوم موجودًا.

الشكل (١٩-١)

الحالات وانتقال الحالات في المثال ١٩-١

الحالة
State الحالة ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑬ ⑭ ⑮ ⑯ ⑰ ⑱ ⑲ ⑳
m m m r r r r r r m r r m r r m r m m r
State transition تحول الحالة ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑬ ⑭ ⑮ ⑯ ⑰ ⑱ ⑲ ⑳

المثال ١٩-١:

نظام له حالتان: سوء استخدام (m) واستخدام عادي (r). تمّ رصد وجود سلسلة لحالات النظام لغرض استكشاف نموذج سلسلة ماركوف: *mmmrrrrrrmrrmrrmrmmr*. قم ببناء نموذج سلسلة ماركوف باستخدام السلسلة المرصودة من حالات النظام، واحسب احتمال توليد سلسلة حالات النظام *mmrmrr*، بواسطة نموذج سلسلة ماركوف. ويبين الشكل ١٩-١ الحالات وانتقال الحالات في السلسلة الاستكشافية المرصودة لحالات النظام. باستخدام المعادلة ١٩-٦ والسلسلة الاستكشافية لحالات النظام *mmmrrrrrrmrrmrrmrmmr*، فإننا نتعلم احتمالات انتقال الحالة التالية:

$$P(m|m) = \frac{N_{mm}}{N_{m}} = \frac{3}{8},$$

لأن انتقالات الحالة ١ و٢ و١٨ هي تحول الحالة $m \to m$، وتحول الحالات ١، ٢، ٣، ١٠، ١٣، ١٦، ١٨، و١٩ هي التحول من الحالة m إلى← أي حالة:

$$P(r|m) = \frac{N_{rm}}{N_{.m}} = \frac{5}{8},$$

لأن انتقالات الحالة ٣، ١٠، ١٣، ١٦، و١٩ هي تحول الحالة من $m \to m$، انتقالات الحالة ١، ٢، ٣، ١٠، ١٣، ١٦، ١٨، و١٩ هي التحول من الحالة m إلى← أي حالة:

$$P(m|r) = \frac{N_{mr}}{N_{.r}} = \frac{4}{11},$$

لأن تحول الحالات ٩، ١٢، ١٥، و١٧ هي انتقال الحالة من $m \to r$، وتحول الحالات ٤، ٥، ٦، ٧، ٨، ٩، ١١، ١٢، ١٤، ١٥، و١٧ هي التحول من الحالة r ← أي حالة:

$$P(r|r) = \frac{N_{rr}}{N_{.r}} = \frac{7}{11},$$

لأن تحول الحالات ٤، ٥، ٦، ٧، ٨، ١١، و١٤ هو التحول من الحالة $m \to m$، وتحول الحالات ٤، ٥، ٦، ٧، ٨، ٩، ١١، ١٢، ١٤، ١٥، و١٧ هي التحول من الحالة m ← أي حالة.

باستخدام المعادلة ١٩-٧ السلسلة الاستكشافية للحالات $mmmrrrrrmrrmrrmrmmr$، فإننا نتعلم الاحتمالات المبدئية للحالة التالية:

$$P(m) = \frac{N_m}{N} = \frac{8}{20},$$

لأن الحالات ١، ٢، ٣، ١٠، ١٣، ١٦، ١٨، و١٩ هي الحالة m، وهناك ٢٠ حالة في سلسلة الحالات:

$$P(r) = \frac{N_r}{N} = \frac{12}{20},$$

لأن الحالات ٤، ٥، ٦، ٧، ٨، ٩، ١١، ١٢، ١٤، ١٥، ١٧، ٢٠ هي الحالة r، وهناك ٢٠ حالة في سلسلة الحالات. وبعد تعلم جميع المعلمات في نموذج سلسلة ماركوف، نقوم بحساب احتمال أن النموذج يولد سلسلة الحالات: *mmrmrr*.

$$P(mmrmrr) = P(s_1)P(s_2|s_1)P(s_3|s_2)P(s_4|s_3)P(s_5|s_4)P(s_6|s_5)$$

$$= P(m)P(m|m)P(r|m)P(m|r)P(r|m)P(r|r)$$

$$= \left(\frac{8}{20}\right)\left(\frac{3}{8}\right)\left(\frac{5}{8}\right)\left(\frac{4}{11}\right)\left(\frac{5}{8}\right)\left(\frac{7}{11}\right) = 0.014.$$

١٩-٢ نماذج ماركوف المخفية (Hidden Markov Models):

في نموذج ماركوف المخفي، يتم مراقبة ورصد ملحوظة البيانات x في كل مرحلة، ولكن الحالة s في كل مرحلة فإنها غير مرصودة. على الرغم من عدم رصد الحالة في كل مرحلة، فإن تسلسل ملحوظات البيانات المرصودة هو نتيجة لتحولات الحالة وظهور ملحوظة بيانات مرصودة من الحالات لدى وصولها في كل حالة. بالإضافة إلى الاحتمالات المبدئية للحالة واحتمالات تحول الحالة، يتم أيضًا تعريف احتمال ظهور x من كل حالة s، $P(x|s)$، كاحتمال الظهور *(emission probability)* في نموذج ماركوف المخفي.

$$\sum_x P(x|s) = 1. \qquad (١٩-٨)$$

يتم افتراض أن ملحوظات البيانات المرصودة مستقلة عن بعضها البعض، وأن احتمال ظهور x من كل حالة s لا يعتمد على الحالات الأخرى.

ويتم استخدام نموذج ماركوف المخفي لتحديد احتمال توليد سلسلة معطاه من الملحوظات المرصودة، $x_1, \ldots, x_N$ في المراحل، N ,... ,1،بواسطة نموذج ماركوف المخفي. باستخدام أي طريقة من طرق المسار (*path method*) (*Theodoridis and Koutroumbas, 1999*)، يتم حساب هذا الاحتمال على النحو التالي:

$$\sum_{i=1}^{S^N} P\left(x_1, \ldots, x_N | s_{1_i}, \ldots, s_{N_i}\right) P\left(s_{1_i}, \ldots, s_{N_i}\right) = \sum_{i=1}^{S^N} P\left(s_{1_i}\right) P\left(x_1 | s_{1_i}\right) \prod_{n=2}^{N} P\left(s_{n_i} | s_{n-1_i}\right) P\left(x_n | s_{n_i}\right), \qquad (٩-١٩)$$

حيث إن:

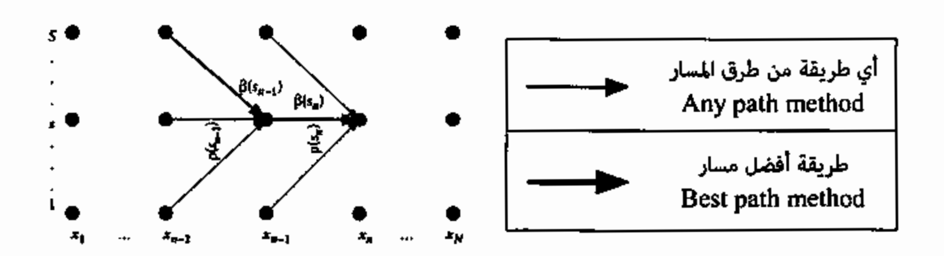
i هو مؤشر لسلسلة الحالات الممكنة، s_{1_i} ،...،s_{N_i}، وهناك عدد S^N من سلاسل الحالات الممكنة، بشكل كامل.

$P(s_{1_i})$ هو الاحتمال الأوّلي للحالة، $P\ (s_{n_i} |\ s_{n-1_i})$ هو احتمال تحول الحالة

$P(x_n \mid s_{n_i})$ هو احتمال الظهور

الشكل (١٩-٢)

أي طريقة من طرق المسار وطريقة المسار الأفضل لنماذج ماركوف المخفية



يبين الشكل ١٩-٢ المراحل، N ,... ,1، والحالات، S ,...,1، والملحوظات المرصودة في المراحل، $x_1, \ldots, x_N$ اللازمة في حساب المعادلة ١٩-٩. لتنفيذ الحسابات في المعادلة ١٩-٩، نقوم بتعريف $\rho(s_N)$ على أنه احتمال أن الحالة (١) يتم الوصول للحالة s_n في المرحلة n،

و(٢) تمّ إظهار الملحوظات المرصودة $x_1, \ldots, x_{n-1}$ في المراحل من 1 إلى $n-1$، و(٣) تمّ إظهار الملحوظة المرصودة x_n من الحالة s_{n_i} في المرحلة n. يمكن حساب $\rho(s_n)$ بشكل تكراري على النحو التالي:

$$\rho(s_n) = \sum_{s_{n-1}=1}^{S} \rho(s_{n-1}) P(s_n|s_{n-1})\, P(x_n|s_n), \qquad (١٠-١٩)$$

$$\rho(s_1) = P(s_1)P(x_1|s_1). \qquad (١١-١٩)$$

وهو ما يعني، $\rho(S_n)$ يمثـــل مجمـــوع احتمالات أن البدء من كل حالة ممكنة $S_n = 1, \ldots, S$ في المرحلة $n-1$ مع $x_1, \ldots, x_{n-1}$ قد ظهرت بالفعل، ونتحول إلى الحالة S_n في المرحلة n التي تُظهِر x_n كما هو موضح في الشكل ١٩-٢. باستخدام المعادلات ١٩-١٠ و١٩-١١، يمكن حساب المعادلة ١٩-٩ على النحو التالي:

$$\sum_{i=1}^{S^N} P\left(x_1, \ldots, x_N | s_{1_i}, \ldots, s_{N_i}\right) P\left(s_{1_i}, \ldots, s_{N_i}\right) = \sum_{s_N=1}^{S} \rho(S_N). \qquad (١٢-١٩)$$

وبالتالي، باستخدام أي طريقة من طرق المسار، يتم استخدام المعادلات من ١٩-١٠ إلى ١٩-١٢ لحساب احتمال أن يقوم نموذج ماركوف المخفي بتوليد سلسلة من الملحوظات المرصودة، $x_1, \ldots, x_N$ تبدأ أي طريقة من طرق المسار بحساب جميع $\rho(s_1)$ لـ $s_1 = 1, \ldots, S$ باستخدام المعادلة ١٩-١١، ثم يستخدم $\rho(s_1)$ لحساب جميع $\rho(s_2)$ حيث $s_2 = 1, \ldots, S$ باستخدام المعادلة ١٩-١٠، ويستمر ذلك على طول الطريق للحصول على جميع $\rho(s_N)$ لـ $s_N = 1, \ldots, S$، والتي يتم استخدامها في النهاية في المعادلة ١٩-١٢ لإكمال العملية الحسابية.

إن التكلفة الحاسوبية لإجراء طريقة من طرق المسار تُعتبَر مرتفعة، لأن كل سلاسل/مسارات الحالة الممكنة التي عددها S^N من سـلاسـل أو مسارات الحالة من المرحلة ١ إلى المرحلة N تُسهِم في العملية الحسابية. بـدلًا من استخـدام المعـادلة ١٩-٩، فإنَّ أفضل

طريقة مسار تستخدم المعـادلة ١٩-١٣ لحســاب احتمال توليـد ســلســلة معطاة من الملحوظات المرصودة، $x_1, ..., x_N$ في المراحل، $1, ..., N$، بواسطة نموذج ماركوف المخفي:

$$\max_{i=1}^{S^N} P(x_1, \dots, x_N | s_{i_1}, \dots, s_{i_N}) P(s_{i_1}, \dots, s_{i_N}) = \max_{i=1}^{S^N} P(s_{i_1}) P(x_1 | s_{i_1}) \prod_{n=2}^{N} P(s_{i_n} | s_{i_{n-1}}) P(x_n | s_{i_n}). \quad (١٩\text{-}١٣)$$

وهو ما يعني، بدلًا من إجراء عملية على مستوى كل سلاسل الحالة الممكنة في المعادلة ١٩-٩ لأي طريقة من طرق المسار، فإنَّ أفضل طريقة مسار تستخدم الحد الأقصى لاحتمال توليد سلسلة من الملحوظات المرصودة، $x_1, ..., x_N$ من قبل أي سلسلة ممكنة للحالة من المرحلة ١ إلى المرحلة N. نقوم بتعريف $\beta(S_n)$ على انها احتمال أن (١) يتم الوصول إلى الحالة S_n في المرحلةn من خلال أفضل مسار، (٢) تَظهَر الملحوظات المرصودة $x_1, ..., x_{n-1}$ في المراحل 1 إلى المرحلة $n - 1$، و(٣) تَظهَر الملحوظة المرصودة x_n من الحالة S_n في المرحلة n. يمكن حساب $\beta(S_n)$ بشكل تكراري كما يلي باستخدام مبدأ بيلمان (*Bellman's principle*) (ثيودوريديس وكوترومباس، ١٩٩٩):

$$\beta(s_n) = \max_{s_{n-1}=1}^{S} [\beta(s_{n-1}) P(s_n | s_{n-1}) P(x_n | s_n)] \quad (١٩\text{-}١٤)$$

$$\beta(s_1) = P(s_1) P(x_1 | s_1). \quad (١٩\text{-}١٥)$$

يتم حساب المعادلة ١٩-١٣ باستخدام المعادلة ١٩-١٦:

$$\max_{i=1}^{S^N} P(x_1, \dots, x_N | s_{i_1}, \dots, s_{i_N}) P(s_{i_1}, \dots, s_{i_N}) = \max_{S_N=1}^{S} \beta(s_N). \quad (١٩\text{-}١٦)$$

وتُستخدَم خوارزمية فيتربي (*Viterbi algorithm*)، (*Viterbi, 1967*)، على نطاق واسع، لحساب التحويل اللوغاريتمي للمعادلات من ١٩-١٣ إلى ١٩-١٦.

تتطلب طريقة أفضل مسار أقل تكلفة حاسوبية لتخزين وحساب الاحتمالات بشكل أكثر من أي طريقة مسار أخرى لأن التكلفة الحاسوبية في أي مرحلة n تستلزم فقط أفضل S من المسارات. بالرغم من ذلك، بالمقارنة مع أي طريقة من طرق المسار، فإن أفضل طريقة مسار هي طريقة البديل الأمثل الفرعي لحساب احتمال توليد سلسلة معطاة من الملحوظات المرصودة، $x_1, ..., x_N$ في المراحل، $1, ..., N$، بواسطة نموذج ماركوف المخفي، فقط لأنه يتم استخدام أفضل مسار بدلًا من كل المسارات الممكنة لتحديد احتمال رصد $x_1, ..., x_N$ علمًا بأن كل المسارات الممكنة في نموذج ماركوف المخفي من الممكن أن تولّد سلسلة للملحوظات المرصودة.

يتم استخدام نماذج ماركوف المخفية على نطاق واسع في التعرف على السرعة (*speed recognitien*)، والتعرف على الحروف المكتوبة بخط اليد، ومعالجة اللغة الطبيعية، والتعرف على تسلسل الحمض النووي، وهلُم جرا. من خلال تطبيق نماذج ماركوف المخفية في التعرف على الأرقام (*digits*) المكتوبة بخط اليد (*Bishop, 2006*) وهي: $0,1,..., 9$، يتم بناء نموذج ماركوف المخفي لكل رقم. يتم اعتبار أن كل رقم لديه سلسلة من مسارات الخط، $x_1, ..., x_N$ في المراحل $1, ..., N$. كل نموذج من نماذج ماركوف الخفية يكون لديه ١٦ من الحالات الكامنة (*latent state*)، كل منها يمكنه أن يظهر أو ينبعث منه خط مقطع ذو طول ثابت مع زاوية واحدة من ١٦ زاوية ممكنة. وبالتالي، يمكن تحديد توزيع الظهور هذا بمصفوفة 16×16 مع احتمال ظهور أي من الـ ١٦ زاوية من كل من الـ ١٦ حالة. يتم تدريب نموذج ماركوف المخفي لكل رقم لتحديد التوزيع الأولي للاحتمالات، ومصفوفة احتمال التحول، واحتمالات الظهور باستخدام ٤٥ مثال من الأمثلة المكتوبة بخط اليد للأرقام. إذا كان لدينا رقم مكتوب بخط اليد للتعرف عليه، يتم حساب احتمال أن يتم توليد الرقم المكتوب بخط اليد من قبل نموذج ماركوف المخفي لكل رقم. يتم تصنيف الأرقام المكتوبة بخط اليد على أنها الأرقام التي نموذج ماركوف المخفي لها ينتج أعلى احتمال لتوليد الأرقام المكتوب بخط اليد.

وبالتالي، لتطبيق نماذج ماركوف المخفية على مشكلة التصنيف، يتم بناء نموذج ماركوف المخفي لكل فئة من الفئات المستهدفة. بإعطاء سلسلة ملحوظات مرصودة، يتم حساب احتمال توليد سلسلة الملحوظات المرصودة هذه من قبل كل نموذج من نماذج ماركوف المخفية باستخدام أي طريقة مسار أو أفضل طريقة مسار. يتم تصنيف سلسلة الملحوظات

المرصودة المعطاة إلى الفئة المستهدفة التي نموذج ماركوف المخفي لها ينتج أعلى احتمال لتوليد سلسلة الملحوظات المرصودة.

٣-١٩ تعلم نماذج ماركوف المخفية (Learning Hidden Markov Models):

تتضمن مجموعة معلَمات النموذج لنموذج ماركوف المخفي، A، احتمالات تحول الحالة، $P(j|i)$، والاحتمالات الأولية للحالة، $P(i)$، واحتمالات الظهور، $P(x|i)$:

$$A = \{P(j|i), P(i), P(x|i)\}. \quad (١٧\text{-}١٩)$$

هناك حاجة لتعلُم معلمات النموذج من مجموعة البيانات التدريبية التي تحتوي على سلسلة N من الملحوظات المرصودة، $X=x_1, \ldots, x_n$ بما أن الحالات (*states*) لا يمكن ملاحظتها مباشرةً، فإنه لا يمكن استخدام المعادلات ١٩-٦ و١٩-٧ لمعرفة معلمات النموذج مثل احتمالات تحول الحالة، والاحتمالات الأولية للحالة. بدلًا من ذلك، يتم استخدام طريقة تضخيم التوقع (*Expectation Maximization - EM*) لتقدير معلمات النموذج، التي تقوم بتضخيم احتمال الحصول على سلسلة الملحوظات المرصودة من النموذج الذي له معلمات نموذج مُقدرة، $P(X|A)$. الخطوات التالية توضح طريقة تضخيم التوقع (*EM*):

١- إسناد القيم الأولية لمعلمات النموذج، A، واستخدام هذه القيم لحساب $P(X \mid A)$.

٢- إعادة تقدير معلمات النموذج للحصول على $\hat{A}$، وحساب $P(X \mid \hat{A})$.

٣- إذا كان $\in\ < P(X \mid \hat{A}) - P(X \mid A)$، ليكن $A = \hat{A}$ لأن $\hat{A}$ تُحسِّن من احتمال الحصول على سلسلة الملحوظات المرصودة من $\hat{A}$ أكثر من A، وانتقل إلى الخطوة ٢؛ وخلاف ذلك، توقف لأن $P(\hat{A})$ هي أسوأ من أو تشابه $P(A)$، وخذ A على أنها مجموعة نهائية من معلمات النموذج.

في الخطوة ٣، $\in$ هو الحد (*threshold*) المُحدَد مسبقًا لتحسين احتمال توليد سلسلة الملحوظات المرصودة X من معلمات النموذج.

يتم حساب $P(x|\hat{A})$ و$P(x|A)$ في طريقة تضخيم التوقع (*EM*) المذكورة أعلاه باستخدام المعادلة ١٩-١٢ لأي طريقة مسار، وتُستخدَم المعادلة ١٩-١٦ للحصول على أفضل طريقة مسار. إذا كانت ملحوظةُ البيانات المرصودة منفصلةً (*discrete*)، وبالتالي سلسلة الملحوظات هي عضو في مجموعة محدودة من سلاسل الملحوظات، يتم استخدام طريقة إعادة التقدير باوم- ولش (*Baum-Welch*) لإعادة تقدير- معلمات النموذج في الخطوة ٢ من طريقة تضخيم التوقع (*EM*) المذكورة آنفًا. يصف ثيودوريديس وكوترومباس، (*Theodoridis and Koutroumbas, 1999*)، طريقة باوم- ولش لإعادة التقدير على النحو التالي. لتكن $\theta_n(i, j, X \mid A)$ هي الاحتمال أن (١) يمر المسار من خلال الحالة i في المرحلة n، (٢) يمر المسار من خلال الحالة j في المرحلة اللاحقة $n+1$، و(٣) ويقوم النموذج بتوليد سلسلة الملاحظات X باستخدام نموذج المعلمات A. لتكن $\varphi_n(i,X|A)$ هي احتمال أن (١) يمر المسار من خلال الحالة i في المرحلة n، و(٢) ويقوم النموذج بتوليد سلسلة الملحوظات X باستخدام معلمات النموذج A. لتكن $\omega_n(i)$ هي الاحتمال أن يكون لدينا الملحوظات $x_{n+1}, \ldots, x_N$ في المراحل $n+1, \ldots, N$، علمًا بأن المسار يمر من خلال i في المرحلة n. بالنسبة لأي طريقة للمسار، يمكن حساب $\omega_n(i)$ بشكل تكراري لـ $n = N - 1, \ldots, 1$ على النحو التالي:

$$\omega_n(i) = P(x_{n+1}, \ldots, x_N | s_n = i, A) = \sum_{s_{n+1}=1}^{S} \omega_{n+1}(s_{n+1}) P(s_{n+1}|s_n = i) P(x_{n+1}|s_{n+1})$$

(١٨-١٩)

$$\omega_N(i) = 1, \;\; i = 1, \ldots, S. \qquad (١٩\text{-}١٩)$$

للحصول على أفضل طريقة للمسار، يمكن حساب $\omega_n(i)$ بشكل تكراري لـ $n = N - 1, \ldots, 1$،على النحو التالي:

$$\omega_n(i) = P(x_{n+1}, \ldots, x_N | s_n = i, A) = \max_{s_{n+1}=1}^{S} \omega_{n+1}(s_{n+1}) P(s_{n+1}|s_n = i) P(x_{n+1}|s_{n+1})$$

(٢٠-١٩)

$$\omega_N(i) = 1, \ \ i = 1, \dots, S. \qquad (٢١-١٩)$$

يكون لدينا أيضًا:

$$\varphi_n(i, X|A) = \rho_n(i)\omega_n(i), \qquad (٢٢-١٩)$$

حيث تدل $\rho_n(i)$ على $\rho(s_n = i)$، والتي يتم حسابها باستخدام المعادلات ١٩-١٠ و١٩-١١. معلمة النموذج $P(i)$ هي العدد المتوقع من المرات التي تحدث فيها الحالة i في المرحلة ١، إذا كان لدينا سلسلة الملحوظات X ومعلمات النموذج A، وهو ما يعني، $P(i \mid X, A)$. معلمة النموذج $P(j|i)$ هي عدد المرات المتوقعة التي يحدث فيها التحول من الحالة i للحالة j، إذا كان لدينا سلسلة الملحوظات X ، ومعلمات النموذج A، وهو ما يعني، $P(i, j|X, A) \,/\, P(i|X, A)$. يتم إعادة تقدير معلمات النموذج على النحو التالي:

$$\hat{P}(i) = P(i|X, A) = \frac{\varphi_1(i, X|A)}{P(X|A)} = \frac{\rho_1(i)\omega_1(i)}{P(X|A)} \qquad (٢٣-١٩)$$

$$\hat{P}(j|i) \ = \frac{P(i, j|X, A)}{P(i|X, A)} = \frac{\sum_{n=1}^{N-1} \theta_n(i, j, X|A)/P(X|A)}{\sum_{n=1}^{N-1} \varphi_n(i, X|A)/P(X|A)}$$

$$= \frac{\sum_{n=1}^{N-1} \rho_n(i)P(j|i)P(x_{n+1}|j)\omega_{n+1}(j)/P(X|A)}{\sum_{n=1}^{N-1} \rho_n(i)\omega_n(i) \,/P(X|A)}$$

$$= \frac{\sum_{n=1}^{N-1} \rho_n(i)P(j|i)P(x_{n+1}|j)\omega_{n+1}(j)}{\sum_{n=1}^{N-1} \rho_n(i)\omega_n(i)} \qquad (٢٤-١٩)$$

$$\hat{P}(x=v|i)=\frac{\sum_{n=1}^{N}\varphi_{n\&x=v}(i)/P(X|A)}{\sum_{n=1}^{N}\varphi_{n}(i)/P(X|A)}=\frac{\sum_{n=1}^{N}\varphi_{n\&x_n=v}(i)}{\sum_{n=1}^{N}\varphi_{n}(i)}$$

$$=\frac{\sum_{n=1}^{N}\rho_{n\&x=v}(i)\omega_{n\&x_n=v}(i)}{\sum_{n=1}^{N}\rho_{n}(i)\omega_{n}(i)} \qquad (\text{٢٥-١٩})$$

حيث:

$$\varphi_{n\&x_n=v}(i)=\begin{cases}\varphi_n(i) & \text{if } x_n=v\\ 0 & \text{if } x_n\neq v\end{cases}, \qquad (\text{٢٦-١٩})$$

$$\rho_{n\&x_n=v}(i)=\begin{cases}\rho_n(i) & \text{if } x_n=v\\ 0 & \text{if } x_n\neq v\end{cases}, \qquad (\text{٢٧-١٩})$$

$$\omega_{n\&x_n=v}(i)=\begin{cases}\omega_n(i) & \text{if } x_n=v\\ 0 & \text{if } x_n\neq v\end{cases}, \qquad (\text{٢٨-١٩})$$

وv هي أحد متجهات القيم المنفصلة التي قد تأخذها x

المثال ١٩-٢:

نظام لديه حالتان: سوء الاستخدام (m) والاستخدام المنتظم (r)، يمكن لكل منهما أن ينتج واحدًا من ثلاثة أحداث: F، G، وH، ويتم رصد سلسلة من خمسة أحداث: $FFFHG$ باستخدام أي من طرق المسار، قم بتنفيذ تكرار واحد من إعادة تقدير معلمات النموذج في طريقة تضخيم التوقع (EM) لتعلُم واستكشاف نموذج ماركوف مخفي من السلسلة المرصودة للأحداث. في الخطوة ١ من طريقة تضخيم التوقع (EM)، يتم إسناد القيم العشوائية التالية لمعلمات النموذج بشكل مبدئي:

$$P(m)=0.4 \quad P(r)=0.6$$

$$P(m|m)=0.375 \quad P(r|m)=0.625 \quad P(m|r)=0.364 \quad P(r|r)=0.636$$

$$P(F|m) = 0.7 \quad P(G|m) = 0.1 \quad P(H|m) = 0.2$$

$$P(F|r) = 0.3 \quad P(G|r) = 0.4 \quad P(H|r) = 0.4.$$

باستخدام هذه المعلمات للنموذج، نقوم بحساب $P(X = FFFHG \mid A)$ باستخدام المعادلات ١٩-١٠، ١٩-١١ و١٩-١٢ لأي طريقة مسار:

$$\rho_1(m) = \rho(s_1 = m) = P(s_1 = m)P(x_1 = F|s_1 = m) = (0.4)(0.7) = 0.28$$

$$\rho_1(r) = \rho(s_1 = r) = P(s_1 = r)P(x_1 = F|s_1 = r) = (0.6)(0.2) = 0.12$$

$$\rho_2(m) = \rho(s_2 = m) = \sum_{s_1=1}^{2} \rho(s_1)P(s_2|s_1)\,P(x_2|s_2)$$

$$= \rho(s_1 = m)P(s_2 = m|s_1 = m)P(x_2 = F|s_2 = m)$$

$$+ \rho(s_1 = r)P(s_2 = m|s_1 = r)P(x_2 = F|s_2 = m)$$

$$= (0.28)(0.375)(0.7) + (0.12)(0.364)(0.7) = 0.1060$$

$$\rho_2(r) = \rho(s_2 = r) = \sum_{s_1=1}^{2} \rho(s_1)P(s_2|s_1)\,P(x_2|s_2)$$

$$= \rho(s_1 = m)P(s_2 = r|s_1 = m)P(x_2 = F|s_2 = r)$$

$$+ \rho(s_1 = r)P(s_2 = r|s_1 = r)P(x_2 = F|s_2 = r)$$

$$= (0.28)(0.625)(0.3) + (0.12)(0.636)(0.3) = 0.0754$$

$$\rho_3(m) = \rho(s_3 = m) = \sum_{s_2=1}^{2} \rho(s_1)P(s_3|s_2)\,P(x_3|s_3)$$

$$= \rho(s_2 = m)P(s_3 = m|s_2 = m)P(x_3 = F|s_3 = m)$$

$$+\rho(s_2 = r)P(s_3 = m|s_2 = r)P(x_3 = F|s_3 = m)$$

$$= (0.1060)(0.375)(0.7) + (0.0754)(0.364)(0.7) = 0.0470$$

$$\rho_3(r) = \rho(s_3 = r) = \sum_{s_2=1}^{2} \rho(s_2)P(s_3|s_2)\,P(x_3|s_3)$$

$$= \rho(s_2 = m)P(s_3 = r|s_2 = m)P(x_3 = F|s_3 = r)$$

$$+\rho(s_2 = r)P(s_3 = r|s_2 = r)P(x_3 = F|s_3 = r)$$

$$= (0.1060)(0.625)(0.2) + (0.0754)(0.636)(0.2) = 0.0228$$

$$\rho_4(m) = \rho(s_4 = m) = \sum_{s_3=1}^{2} \rho(s_3)P(s_4|s_3)\,P(x_4|s_4)$$

$$= \rho(s_3 = m)P(s_4 = m|s_3 = m)P(x_4 = H|s_4 = m)$$

$$+\rho(s_3 = r)P(s_4 = m|s_3 = r)P(x_4 = H|s_4 = m)$$

$$= (0.0470)(0.375)(0.2) + (0.0228)(0.364)(0.2) = 0.0052$$

$$\rho_4(r) = \rho(s_4 = r) = \sum_{s_3=1}^{2} \rho(s_3)P(s_4|s_3)\,P(x_4|s_4)$$

$$= \rho(s_3 = m)P(s_4 = r|s_3 = m)P(x_4 = H|s_4 = r)$$

$$+\rho(s_3 = r)P(s_4 = r|s_3 = r)P(x_4 = H|s_4 = r)$$

$$= (0.0470)(0.625)(0.4) + (0.0228)(0.636)(0.4) = 0.0176$$

$$\rho_5(m) = \rho(s_5 = m) = \sum_{s_4=1}^{2} \rho(s_4)P(s_5|s_4)\,P(x_5|s_5)$$

$$= \rho(s_4 = m)P(s_5 = m|s_4 = m)P(x_5 = G|s_5 = m)$$

$$+ \rho(s_4 = r)P(s_5 = m|s_4 = r)P(x_5 = G|s_5 = m)$$

$$= (0.0052)(0.375)(0.1) + (0.0176)(0.364)(0.1) = 0.0008$$

$$\rho_5(r) = \rho(s_5 = r) = \sum_{s_4=1}^{2} \rho(s_4)P(s_5|s_4)\,P(x_5|s_5)$$

$$= \rho(s_4 = m)P(s_5 = r|s_4 = m)P(x_5 = G|s_5 = r)$$

$$+ \rho(s_4 = r)P(s_5 = r|s_4 = r)P(x_5 = G|s_5 = r)$$

$$= (0.0052)(0.625)(0.4) + (0.0176)(0.636)(0.4) = 0.0058$$

$$P(X = FFFHG|A) = \sum_{s_5=1}^{2} \rho(s_5) = \rho(s_5 = m)\rho(s_5 = r) = 0.0008 + 0.0058$$
$$= 0.0066$$

في الخطوة ٢ من طريقة التوقع (EM)، نقوم باستخدام المعادلات ١٩-٢٣ و١٩-٢٥ لإعادة تقدير معلمات النموذج. نحتاج أولًا إلى استخدام المعادلات ١٩-١٨ و١٩-١٩ لحساب $\omega_n(i)$ $n=5,4,3,2,1$, والتي يتم استخدامها في المعادلات من ١٩-٢٣ إلى ١٩-٢٥:

$$\omega_5(m) = 1 \quad \omega_5(r) = 1$$

$$\omega_4(m) = P(x_5 = G|s_4 = m, A) = \sum_{s_5=1}^{2} \omega_5(s_5)P(s_5|s_4 = m)\,P(x_5 = G|s_5)$$

$$= \omega_5(m)P(s_5 = m|s_4 = m)P(x_5 = G|s_5 = m)$$

$$+ \omega_5(r)P(s_5 = r|s_4 = m)P(x_5 = G|s_5 = r)$$

$$= (1)(0.375)(0.1) + (1)(0.625)(0.4) = 0.2875$$

$$\omega_4(r) = P(x_5 = G|s_4 = r, A) = \sum_{s_5=1}^{2} \omega_5(s_5)P(s_5|s_4 = r)\, P(x_5 = G|s_5)$$

$$= \omega_5(m)P(s_5 = m|s_4 = r)P(x_5 = G|s_5 = m)$$

$$+ \omega_5(r)P(s_5 = r|s_4 = r)P(x_5 = G|s_5 = r)$$

$$= (1)(0.364)(0.1) + (1)(0.636)(0.4) = 0.2908$$

$$\omega_3(m) = P(x_4 = H, x_5 = G|s_3 = m, A) = \sum_{s_4=1}^{2} \omega_4(s_4)P(s_4|s_3 = m)\, P(x_4 = H|s_4)$$

$$= \omega_4(m)P(s_4 = m|s_3 = m)P(x_4 = H|s_4 = m)$$

$$+ \omega_4(r)P(s_4 = r|s_3 = m)P(x_4 = H|s_4 = r)$$

$$= (0.2875)(0.375)(0.2) + (0.2908)(0.625)(0.4)$$

$$= 0.0943$$

$$\omega_3(r) = P(x_4 = H, x_5 = G|s_3 = r, A) = \sum_{s_4=1}^{2} \omega_4(s_4)P(s_4|s_3 = r)\, P(x_4 = H|s_4)$$

$$= \omega_4(m)P(s_4 = m|s_3 = r)P(x_4 = H|s_4 = m)$$

$$+ \omega_4(r)P(s_4 = r|s_3 = r)P(x_4 = H|s_4 = r)$$

$$= (0.2875)(0.364)(0.2) + (0.2908)(0.636)(0.4)$$

$$= 0.0949$$

$$\omega_2(m) = P(x_3 = F, x_4 = H, x_5 = G | s_2 = m, A)$$

$$= \sum_{s_3=1}^{2} \omega_3(s_3) P(s_3 | s_2 = m)\, P(x_3 = F | s_3)$$

$$= \omega_3(m) P(s_3 = m | s_2 = m) P(x_3 = F | s_3 = m)$$

$$+ \omega_3(r) P(s_3 = r | s_2 = m) P(x_3 = F | s_3 = r)$$

$$= (0.0943)(0.375)(0.7) + (0.0949)(0.625)(0.2)$$

$$= 0.0366$$

$$\omega_2(r) = P(x_3 = F, x_4 = H, x_5 = G | s_2 = r, A)$$

$$= \sum_{s_3=1}^{2} \omega_3(s_3) P(s_3 | s_2 = r)\, P(x_3 = F | s_3)$$

$$= \omega_3(m) P(s_3 = m | s_2 = r) P(x_3 = F | s_3 = m)$$

$$+ \omega_3(r) P(s_3 = r | s_2 = r) P(x_3 = F | s_3 = r)$$

$$= (0.0943)(0.364)(0.7) + (0.0949)(0.636)(0.2)$$

$$= 0.0361$$

$$\omega_1(m) = P(x_2 = F, x_3 = F, x_4 = H, x_5 = G | s_1 = m, A)$$

$$= \sum_{s_2=1}^{2} \omega_2(s_2) P(s_2|s_1 = m)\, P(x_s = F|s_2)$$

$$= \omega_2(m) P(s_2 = m|s_1 = m) P(x_2 = F|s_2 = m)$$

$$+ \omega_2(r) P(s_2 = r|s_1 = m) P(x_2 = F|s_2 = r)$$

$$= (0.0366)(0.375)(0.7) + (0.0361)(0.625)(0.2)$$

$$= 0.0141$$

$$\omega_1(r) = P(x_2 = F, x_3 = F, x_4 = H, x_5 = G | s_1 = r, A)$$

$$= \sum_{s_2=1}^{2} \omega_2(s_2) P(s_2|s_1 = r)\, P(x_s = F|s_2)$$

$$= \omega_2(m) P(s_2 = m|s_1 = r) P(x_2 = F|s_2 = m)$$

$$+ \omega_2(r) P(s_2 = r|s_1 = r) P(x_2 = F|s_2 = r)$$

$$= (0.0366)(0.364)(0.7) + (0.0361)(0.636)(0.2)$$

$$= 0.0139.$$

نقوم الآن باستخدام المعادلات ١٩-٢٣ و١٩-٢٥ لإعادة تقدير معلمات النموذج:

$$\hat{P}(m) = \frac{\rho_1(m)\omega_1(m)}{P(X = FFFHG|A)} = \frac{(0.28)(0.0141)}{(0.0066)} = 0.5982$$

$$\hat{P}(r) = \frac{\rho_1(r)\omega_1(r)}{P(X = FFFHG|A)} = \frac{(0.12)(0.0139)}{(0.0066)} = 0.2527$$

$$\hat{P}(m|m) = \frac{\sum_{n=1}^{4} \rho_n(m)P(m|m)P(x_{n+1}|m)\omega_{n+1}(m)}{\sum_{n=1}^{4} \rho_n(m)\omega_n(m)}$$

$$= \frac{\begin{bmatrix} \rho_1(m)P(m|m)P(x_2 = F|m)\omega_2(m) \\ + \rho_2(m)P(m|m)P(x_3 = F|m)\omega_3(m) \\ + \rho_3(m)P(m|m)P(x_4 = H|m)\omega_4(m) \\ + \rho_4(m)P(m|m)P(x_5 = G|m)\omega_5(m) \end{bmatrix}}{\begin{bmatrix} \rho_1(m)\omega_1(m) \\ +\rho_2(m)\omega_2(m) \\ +\rho_3(m)\omega_3(m) \\ +\rho_4(m)\omega_4(m) \end{bmatrix}}$$

$$= \frac{\begin{bmatrix} (0.28)(0.375)(0.7)(0.0366) + (0.1060)(0.375)(0.7)(0.0943) \\ +(0.0470)(0.375)(0.2)(0.2875) + (0.0052)(0.375)(0.1)(1) \end{bmatrix}}{[(0.28)(0.0141) + (0.1060)(0.0366) + (0.0470)(0.0943) + (0.0052)(0.2875)]}$$

$$= 0.4742$$

$$\hat{P}(r|m) = \frac{\sum_{n=1}^{4} \rho_n(m)P(r|m)P(x_{n+1}|r)\omega_{n+1}(r)}{\sum_{n=1}^{4} \rho_n(m)\omega_n(m)}$$

$$= \frac{\begin{bmatrix} \rho_1(m)P(r|m)P(x_2 = F|r)\omega_2(r) \\ + \rho_2(m)P(r|m)P(x_3 = F|r)\omega_3(r) \\ + \rho_3(m)P(r|m)P(x_4 = H|r)\omega_4(r) \\ + \rho_4(m)P(r|m)P(x_5 = G|r)\omega_5(r) \end{bmatrix}}{\begin{bmatrix} \rho_1(m)\omega_1(m) \\ +\rho_2(m)\omega_2(m) \\ +\rho_3(m)\omega_3(m) \\ +\rho_4(m)\omega_4(m) \end{bmatrix}}$$

$$= \frac{\begin{bmatrix} (0.28)(0.625)(0.2)(0.0361) + (0.1060)(0.625)(0.2)(0.0949) \\ +(0.0470)(0.625)(0.4)(0.2908) + (0.0052)(0.325)(0.4)(1) \end{bmatrix}}{[(0.28)(0.0141) + (0.1060)(0.0366) + (0.0470)(0.0943) + (0.0052)(0.2875)]}$$

$$= 0.5262$$

$$\hat{P}(m|r) = \frac{\sum_{n=1}^{4} \rho_n(r)P(m|r)P(x_{n+1}|m)\omega_{n+1}(m)}{\sum_{n=1}^{4} \rho_n(r)\omega_n(r)}$$

$$= \frac{\begin{bmatrix} \rho_1(r)P(m|r)P(x_2 = F|m)\omega_2(m) \\ + \rho_2(r)P(m|r)P(x_3 = F|m)\omega_3(m) \\ + \rho_3(r)P(m|r)P(x_4 = H|m)\omega_4(m) \\ + \rho_4(r)P(m|r)P(x_5 = G|m)\omega_5(m) \end{bmatrix}}{\begin{bmatrix} \rho_1(r)\omega_1(r) \\ +\rho_2(r)\omega_2(r) \\ +\rho_3(r)\omega_3(r) \\ +\rho_4(r)\omega_4(r) \end{bmatrix}}$$

$$= \frac{\begin{bmatrix} (0.12)(0.364)(0.7)(0.0366) + (0.0754)(0.364)(0.7)(0.0943) \\ +(0.0228)(0.364)(0.2)(0.2875) + (0.0176)(0.364)(0.1)(1) \end{bmatrix}}{[(0.12)(0.0139) + (0.0754)(0.0361) + (0.0228)(0.0949) + (0.0176)(0.2908)]}$$

$$= 0.3469$$

$$\hat{P}(r|r) = \frac{\sum_{n=1}^{4} \rho_n(r)P(r|r)P(x_{n+1}|r)\omega_{n+1}(r)}{\sum_{n=1}^{4} \rho_n(r)\omega_n(r)}$$

$$= \frac{\begin{bmatrix} \rho_1(r)P(r|r)P(x_2 = F|r)\omega_2(r) \\ + \rho_2(r)P(r|r)P(x_3 = F|r)\omega_3(r) \\ + \rho_3(r)P(r|r)P(x_4 = H|r)\omega_4(r) \\ + \rho_4(r)P(r|r)P(x_5 = G|r)\omega_5(r) \end{bmatrix}}{\begin{bmatrix} \rho_1(r)\omega_1(r) \\ +\rho_2(r)\omega_2(r) \\ +\rho_3(r)\omega_3(r) \\ +\rho_4(r)\omega_4(r) \end{bmatrix}}$$

$$= \frac{\begin{bmatrix} (0.12)(0.636)(0.2)(0.0361) + (0.0754)(0.636)(0.2)(0.0949) \\ +(0.0228)(0.636)(0.4)(0.2908) + (0.0176)(0.636)(0.4)(1) \end{bmatrix}}{[(0.12)(0.0139) + (0.0754)(0.0361) + (0.0228)(0.0949) + (0.0176)(0.2908)]}$$

$$= 0.6533$$

$$\hat{P}(x = F|m) = \frac{\sum_{n=1}^{5} \rho_{n\&x_n=F}(m)\omega_{n\&x_n=F}(m)}{\sum_{n=1}^{5} \rho_n(m)\omega_n(m)}$$

$$= \frac{\begin{matrix} \rho_{1\&x_1=F}(m)\omega_{1\&x_1=F}(m) + \rho_{2\&x_2=F}(m)\omega_{2\&x_2=F}(m) \\ + \rho_{3\&x_3=F}(m)\omega_{3\&x_3=F}(m) + \\ + \rho_{4\&x_4=F}(m)\omega_{4\&x_4=F}(m) + \rho_{5\&x_5=F}(m)\omega_{5\&x_5=F}(m) \end{matrix}}{\rho_1(m)\omega_1(m) + \rho_2(m)\omega_2(m) + \rho_3(m)\omega_3(m) + \rho_4(m)\omega_4(m) + \rho_5(m)\omega_5(m)}$$

$$= \frac{(0.28)(0.0141) + (0.1060)(0.0366) + (0.0470)(0.0943) + (0)(0) + (0)(0)}{(0.12)(0.0141) + (0.1060)(0.0366) + (0.0470)(0.0943) + (0.0052)(0.2875) + (0.0058)(1)}$$

$$= 0.6269$$

$$\hat{P}(x=G|m)=\frac{\sum_{n=1}^{5}\rho_{n\&x_n=G}(m)\omega_{n\&x_n=G}(m)}{\sum_{n=1}^{5}\rho_n(m)\omega_n(m)}$$

$$=\frac{\rho_{1\&x_1=G}(m)\omega_{1\&x_1=G}(m)+\rho_{2\&x_2=G}(m)\omega_{2\&x_2=G}(m)+\rho_{3\&x_3=G}(m)\omega_{3\&x_3=G}(m)+\rho_{4\&x_4=G}(m)\omega_{4\&x_4=G}(m)+\rho_{5\&x_5=G}(m)\omega_{5\&x_5=G}(m)}{\rho_1(m)\omega_1(m)+\rho_2(m)\omega_2(m)+\rho_3(m)\omega_3(m)+\rho_4(m)\omega_4(m)+\rho_5(m)\omega_5(m)}$$

$$=\frac{(0)(0)+(0)(0)+(0)(0)+(0)(0)+(0.0008)(1)}{(0.28)(0.0141)+(0.1060)(0.0366)+(0.0470)(0.0943)+(0.0052)(0.2875)+(0.0008)(1)}$$

$$=0.0550$$

$$\hat{P}(x=H|m)=\frac{\sum_{n=1}^{5}\rho_{n\&x_n=H}(m)\omega_{n\&x_n=H}(m)}{\sum_{n=1}^{5}\rho_n(m)\omega_n(m)}$$

$$=\frac{\rho_{1\&x_1=H}(m)\omega_{1\&x_1=H}(m)+\rho_{2\&x_2=H}(m)\omega_{2\&x_2=H}(m)+\rho_{3\&x_3=H}(m)\omega_{3\&x_3=H}(m)+\rho_{4\&x_4=H}(m)\omega_{4\&x_4=H}(m)+\rho_{5\&x_5=H}(m)\omega_{5\&x_5=H}(m)}{\rho_1(m)\omega_1(m)+\rho_2(m)\omega_2(m)+\rho_3(m)\omega_3(m)+\rho_4(m)\omega_4(m)+\rho_5(m)\omega_5(m)}$$

$$=\frac{(0)(0)+(0)(0)+(0)(0)+(0.0052)(0.2875)+(0)(0)}{(0.28)(0.0141)+(0.1060)(0.0366)+(0.0470)(0.0943)+(0.0052)(0.2875)+(0.0008)(1)}$$

$$=0.1027$$

$$\hat{P}(x=F|r)=\frac{\sum_{n=1}^{5}\rho_{n\&x_n=F}(r)\omega_{n\&x_n=F}(r)}{\sum_{n=1}^{5}\rho_n(r)\omega_n(r)}$$

$$\rho_{1\&x_1=F}(r)\omega_{1\&x_1=F}(r)+\rho_{2\&x_2=F}(r)\omega_{2\&x_2=F}(r)+\rho_{3\&x_3=F}(r)\omega_{3\&x_3=F}(r)+$$

$$= \frac{+\,\rho_{4\&x_4=F}(r)\omega_{4\&x_4=F}(r) + \rho_{5\&x_5=F}(r)\omega_{5\&x_5=F}(r)}{\rho_1(r)\omega_1(r) + \rho_2(r)\omega_2(r) + \rho_3(r)\omega_3(r) + \rho_4(r)\omega_4(r) + \rho_5(r)\omega_5(r)}$$

$$= \frac{(0.12)(0.0139) + (0.0754)(0.0361) + (0.0228)(0.0949) + (0)(0) + (0)(0)}{(0.12)(0.0139) + (0.0754)(0.0361) + (0.0228)(0.0949) + (0.0176)(0.2908) + (0.0058)(1)}$$

$$= 0.3751$$

$$\hat{P}(x = G|r) = \frac{\sum_{n=1}^{5} \rho_{n\&x_n=G}(r)\omega_{n\&x_n=G}(r)}{\sum_{n=1}^{5} \rho_n(r)\omega_n(r)}$$

$$\rho_{1\&x_1=G}(r)\omega_{1\&x_1=G}(r) + \rho_{2\&x_2=G}(r)\omega_{2\&x_2=G}(r) + \rho_{3\&x_3=G}(r)\omega_{3\&x_3=G}(r) +$$

$$= \frac{+\,\rho_{4\&x_4=G}(r)\omega_{4\&x_4=G}(r) + \rho_{5\&x_5=G}(r)\omega_{5\&x_5=G}(r)}{\rho_1(r)\omega_1(r) + \rho_2(r)\omega_2(r) + \rho_3(r)\omega_3(r) + \rho_4(r)\omega_4(r) + \rho_5(r)\omega_5(r)}$$

$$= \frac{(0)(0) + (0)(0) + (0)(0) + (0)(0) + (0.0058)(1)}{(0.12)(0.0139) + (0.0754)(0.0361) + (0.0228)(0.0949) + (0.0176)(0.2908) + (0.0058)(1)}$$

$$= 0.3320$$

$$\hat{P}(x = H|r) = \frac{\sum_{n=1}^{5} \rho_{n\&x_n=H}(r)\omega_{n\&x_n=H}(r)}{\sum_{n=1}^{5} \rho_n(r)\omega_n(r)}$$

$$\rho_{1\&x_1=H}(r)\omega_{1\&x_1=H}(r) + \rho_{2\&x_2=H}(r)\omega_{2\&x_2=H}(r) + \rho_{3\&x_3=H}(r)\omega_{3\&x_3=H}(r) +$$

$$= \frac{+\,\rho_{4\&x_4=H}(r)\omega_{4\&x_4=H}(r) + \rho_{5\&x_5=H}(r)\omega_{5\&x_5=H}(r)}{\rho_1(r)\omega_1(r) + \rho_2(r)\omega_2(r) + \rho_3(r)\omega_3(r) + \rho_4(r)\omega_4(r) + \rho_5(r)\omega_5(r)}$$

$$= \frac{(0)(0) + (0)(0) + (0)(0) + (0.0176)(0.2908) + (0)(0)}{(0.12)(0.0139) + (0.0754)(0.0361) + (0.0228)(0.0949) + (0.0176)(0.2908) + (0.0058)(1)}$$

$$= 0.2929$$

١٩-٤ البرمجيات والتطبيقات (Software and Applications):

تقوم برمجية *HTK* (*Hidden Markov Model Tookit*) (*http://htk.eng.cam.ac.uk*)، بدعم نماذج ماركوف المخفية. قامت يي وزملائها (*Ye, 2008; Ye at al., 2002c, 2004b*) بوصف تطبيق نماذج سلسلة ماركوف للكشف عن الهجوم الإلكتروني. وقام رابينر(*Rabiner, 1989*) بمراجعة تطبيقات نماذج ماركوف المخفية للتعرف على الكلام (*speech recognition*).

التمارين (Exercises):

١-١٩ بالنظر إلى نموذج سلسلة ماركوف في المثال ١٩-١، حدد احتمال رصد سلسلة من حالات النظام: *rmmrmrrmrrrrrrmmm*.

٢-١٩ نظام لديه حالتان، سوء الاستخدام (m) والاستخدام المنتظم (r)، يمكن لكل منها أن ينتج واحدًا من ثلاثة أحداث: *F*، *G*، *A* .*H* لدى نموذج ماركوف المخفي للنظام احتمالات تحول الحالة الأولية، واحتمالات تحول الحالة بالنظر إلى المثال ١٩-١، واحتمالات ظهور الحالة على النحو التالي:

$$P(F|m) = 0.1 \quad P(G|m) = 0.3 \quad P(H|m) = 0.6$$

$$P(F|r) = 0.5 \quad P(G|r) = 0.2 \quad P(H|r) = 0.3.$$

استخدم أي طريقة مسار لتحديد احتمال رصد سلسلة من الأحداث الخمسة: *GHFFH*.

٣-١٩ بالنظر إلى نموذج ماركوف المخفية في التمرين ١٩-٢، قم باستخدام أفضل طريقة لتحديد مسار لتحديد احتمال رصد سلسلة من الأحداث الخمسة: *GHFFH*.

٢٠- تحليل المويجة
Wavelet Analysis

هناك العديد من الأشياء (*objects*) التي لها سلوك دوري وبالتالي تُظهِر سمَـة فريدة في مجال التكرار أو التردد. على سبيل المثال، الأصوات البشرية لها مجموعة من الترددات التي تختلف عن تلك التي لدى بعض الحيوانات. إنَّ الأشياء أو الأجسام في الفضاء، بما في ذلك الأرض تتحرك بتكرارات مختلفة. الأجسام الجديدة في الفضاء يمكن اكتشافها من خلال مراقبة تكرار حركتها الفريدة، والتي تختلف عن تلك الأجسام المعروفة. وبالتالي، فإن سمَـة التكرار أو التردد لأي جسم يمكن أن تكون مفيدة في تحديد الجسم أو الشيء. أن تحليل المويجات (*Wavelet analysis*) يمثل بيانات السلاسل الزمنية في مجال التكرار الزمني (*time-frequency*) باستخدام خصائص البيانات على مر الزمن في تكرارات مختلفة، وبالتالي يسمح لنا بكشف أنماط البيانات الزمنية في تكرارات متنوعة. هناك العديد من أشكال المويجات، على سبيل المثال، هار(*Haar*)، داوبيشيز (*Daubecnies*)، واشتقاق مويجة قوسشيان (*DoG*). في هذا الفصل، نقوم باستخدام مويجة هار (*Haar*) لشرح كيفية عمل تحليل المويجات لتحويل بيانات السلاسل الزمنية إلى بيانات في مجال التكرار الزمن. وترد قائمة من حزم البرمجيات التي تدعم تحليل المويجات. ويتم إعطاء بعض التطبيقات لتحليل المويجات مع المراجع.

٢٠-١ تعريف المويجة (Definition of Wavelet):

يتم تعريف شكل المويجة عن طريق دالتين: دالة القياس (*Scaling Function*) $\varphi(x)$، ودالة المويجة (*Wavelet Function*) $\psi(x)$. تُعدُ دالة القياس لمويجة هار هي دالة خطوة (*Boggess and Narcowich, 2001; Vidakovic, 1999*)، كما هو مبين في الشكل ٢٠-١ :

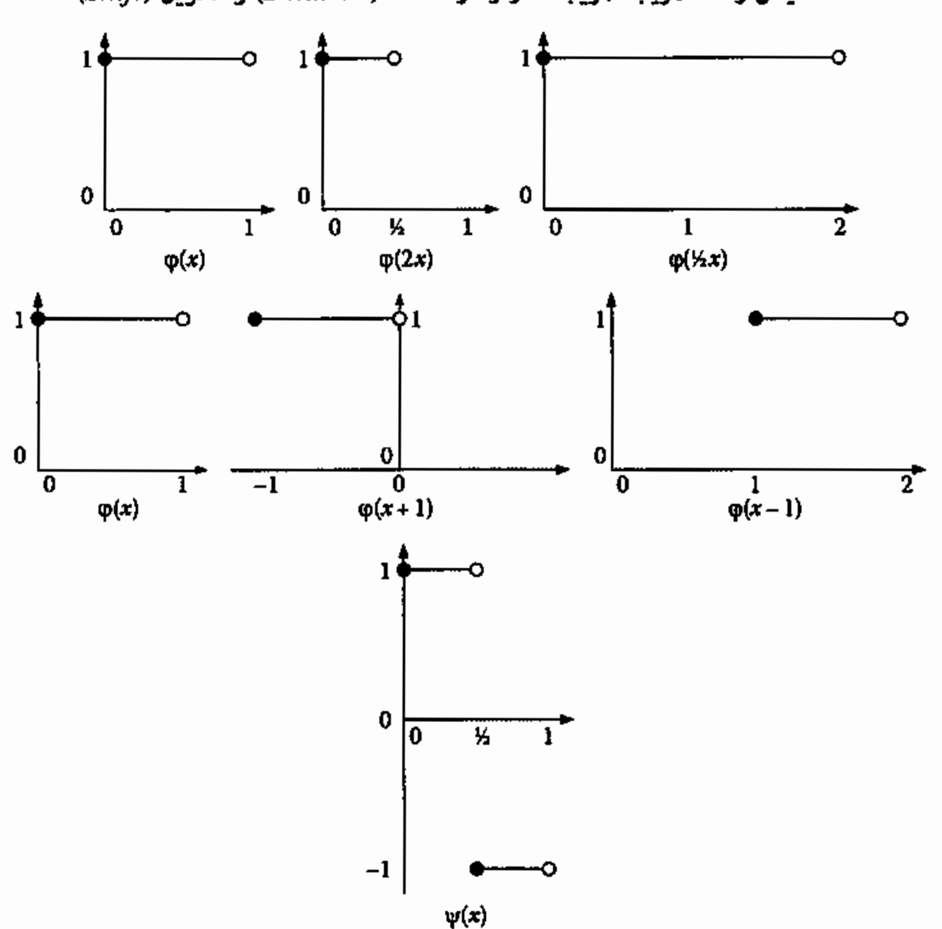
$$\varphi(x) = \begin{cases} 1 & \text{if } 0 \le x < 1 \\ 0 & \text{otherwise} \end{cases}. \qquad (٢٠\text{-}١)$$

يتم تعريف دالة المويجة لمويجة هار (*Haar wavelet*) باستخدام دالة القياس (*Boggess and Narcowich, 2001; Vidakovic, 1999*)، كما هو مبين في الشكل ٢٠-١:

$$\psi(x) = \varphi(2x) - \varphi(2x-1) = \begin{cases} 1 & \text{if } 0 \le x < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} \le x < 1 \end{cases}. \quad (٢٠-٢)$$

الشكل ٢٠-١

دالة القياس ودالة المويجة لمويجة هار وآثار التمدد (*Dilation*) والتحويل (*Shift*)



وبالتالي، فإنَّ دالة المويجة لمويجة هار تمثل التغير في قيمة الدالة من 1 إلى 1- في النطاق $[0,1)$. إن الدالة $\varphi(2x)$ في المعادلة ٢٠-٢ هي دالة الخطوة بارتفاع مقدراه 1 لنطاق قيم x في $[0,\frac{1}{2})$، كما في الشكل ٢٠-١. وبشكل عام، تُعطي المعلَمة a قبل x في $\varphi\ (ax)$ أثرًا تمدديًا على نطاق قيم x مما يعمل على توسيع أو تضييق نطاق x بمقدار $1/a$، كما هو مبين في الشكل ٢٠-١. دالة $\varphi\ (2x-1)$ هي أيضًا دالة خطوة بارتفاع مقدراه 1 لنطاق قيم x في $[\frac{1}{2},1)$. وبشكل عام، فإنَّ المعلَمة b في $\varphi\ (x+b)$ تُعطي أثرًا تحوليًا على نطاق قيم x مما يحرك نطاق x بنسبة b، كما هو مبين في الشكل ٢٠-١. وبالتالي، فإن $\varphi(ax+b)$ تعرف دالة خطوة ذات ارتفاع مقدراه ١ لقيم x في نطـــــاق $[-b/a, (1-b)/a)$، كما هو موضح أدناه، مع الأخذ في الاعتبار أن $a > 0$:

$$0 \leq ax + b < 1$$

$$\frac{-b}{a} \leq x < \frac{1-b}{a}.$$

٢٠-٢ تحويل المويجة لبيانات السلاسل الزمنية

(Wavelet Transform of Time Series Data)

إذا كان لدينا بيانات سلسلة زمنية مع دالة كما هو موضح في الشكل (٢٠—a٢) وعينة من سجلات بيانات عددها ثمانية 0، 2، 0، 0، 6، 8، 6، 8 المأخوذة من هذه الدالة في نقاط الوقت على المحور السيني 0، $\frac{1}{8}$ ، على التوالي، عند الفترات الزمنية $\frac{1}{8}$، $\frac{2}{8}$، $\frac{3}{8}$، $\frac{4}{8}$، $\frac{5}{8}$، $\frac{6}{8}$، $\frac{7}{8}$، أو عند التكرار 8، كما هو موضح في الشكل (٢٠—b٢):

$$a_i,\ \ i = 0, 1, \dots, 2^k - 1, \qquad k = 3 \ \text{or}$$

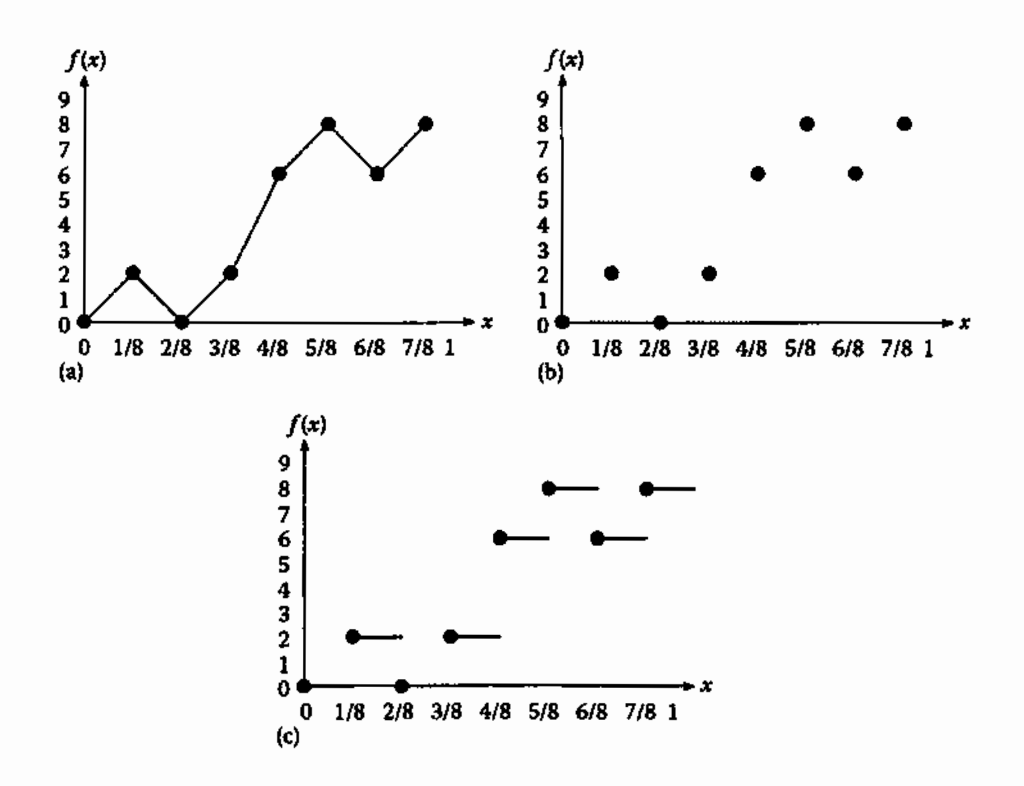
$$a_0 = 0, a_1 = 2, a_2 = 0, a_3 = 2, a_4 = 6, a_5 = 8, a_6 = 6, a_7 = 8,$$

يمكن تقريب الدالة باستخدام عينة سجلات البيانات ودالة القياس لمويجة هار على النحو التالي:

$$f(x) = \sum_{i=0}^{2^k-1} a_i \varphi(2^k x - i) \qquad (٣-٢٠)$$

الشكل (٢٠-٢)

عينة من بيانات سلسلة زمنية من (a) دالة، (b) عينة من سجلات البيانات مأخوذة من الدالة، و(c) تقريب الدالة باستخدام دالة القياس لمويجة هار



$$f(x) = a_0\varphi(2^3x-0) + a_1\varphi(2^3x-1) + a_2\varphi(2^3x-2) + a_3\varphi(2^3x-3) + a_4\varphi(2^3x-4)$$

$$+a_5\varphi(2^3x-5) + a_6\varphi(2^3x-6) + a_7\varphi(2^3x-7)$$

$$f(x) = 0\varphi(2^3x) + 2\varphi(2^3x-1) + 0\varphi(2^3x-2) + 2\varphi(2^3x-3) + 6\varphi(2^3x-4)$$

$$+8\varphi(2^3x-5) + 6\varphi(2^3x-6) + 8\varphi(2^3x-7)$$

في المعادلة ٢٠-٣، فإن $a_i\ \varphi(2^kx-i)$ تُعرف دالة خطوة بارتفاع مقداره a_i، لقيم x في النطاق $[i/2^k, (i+1)/2^k)$. ويبين الشكل (٢٠-٢ c) تقريب الدالة باستخدام دوال الخطوة بارتفاع مقداره يساوي سجلات البيانات الثمانية.

عند الأخذ في الاعتبار أول دالتي خطوة في المعادلة ٢٠-٣، $\varphi(2^kx)$ و$\varphi(2^kx-1)$، واللتان لهما القيمة 1 لقيم x في النطاقين $[0, 1/2^k)$ و$[1/2^k, 2/2^k)$ ، على التوالي، يكون لدينا العلاقات التالية :

$$\varphi(2^{k-1}x) = \varphi(2^kx) + \varphi(2^kx-1) \qquad (٢٠\text{-}٤)$$

$$\psi(2^{k-1}x) = \varphi(2^kx) + \varphi(2^kx-1). \qquad (٢٠\text{-}٥)$$

$\varphi(2^{k-1}x)$ في المعادلة ٢٠-٤ لديها القيمة 1 لقيم x في النطاق $[0, 1/2^{k-1})$، والتي يشمل $[0, 1/2^k)$ و$[1/2^k, 2/2^k)$ معًا. كما تغطي الدالة $\psi(2^{k-1}x)$ في المعادلة ٢٠-٥ أيضًا النطاقين $[0, 1/2^k)$ و$[1/2^k, 2/2^k)$ معًا، ولكن يكون لها القيمة 1 عندما تكون قيم x واقعة في $[0, 1/2^{k-1})$ ويكون للدالة القيمة 1- عندما تكون قيم x واقعة في النطاق $[1/2^k, 2/2^k)$. يتم الحصول على صيغة معادلة مكافئة للمعادلات ٢٠-٤ و٢٠-٥ بإضافة المعادلات ٢٠-٤ و٢٠-٥ وبطرح المعادلة ٢٠-٥ من المعادلة ٢٠-٤ :

$$\varphi(2^kx) = \frac{1}{2}[\varphi(2^{k-1}x) + \psi(2^{k-1}x)] \qquad (٢٠\text{-}٦)$$

$$\varphi(2^kx-1) = \frac{1}{2}[\varphi(2^{k-1}x) - \psi(2^{k-1}x)]. \qquad (٢٠\text{-}٧)$$

في الجانب الأيسر من المعادلات ٢٠-٦ و٢٠-٧، ننظر إلى سجلات البيانات في الفترة الزمنية $1/2^k$ أو التكرار 2^k. في الجانب الأيمن من المعادلات ٢٠-٤ و٢٠-٥، ننظر إلى سجلات البيانات في الفترة الزمنية الأكبر $1/2^{k-1}$ أو التكرار الأقل 2^{k-1}.

وبشكل عام، عند الأخذ في الاعتبار دالتي الخطوة في المعادلة ٢٠-٣، وهما: $\varphi(2^k x - i)$ و$\varphi(2^k x - i - 1)$، واللتان لهما القيمة 1 عندما تكون قيم x واقعة في $[i/2^k, (i + 1)/2^k)$ و$[(i + 1)/2^k, (i + 2)/2^k)$، على التوالي، فإنه يكون لدينا العلاقات التالية:

$$\varphi\left(2^{k-1}x - \frac{i}{2}\right) = \varphi(2^k x - i) + \varphi(2^k x - i - 1) \qquad (٢٠\text{-}٨)$$

$$\psi\left(2^{k-1}x - \frac{i}{2}\right) = \varphi(2^k x - i) - \varphi(2^k x - i - 1) \qquad (٢٠\text{-}٩)$$

$\varphi(2^{k-1}x - i/2)$ في المعادلة ٢٠-٨ يكون لها القيمة 1 عندما تكون قيم x واقعة في النطـــاق $[i/2^k, (i + 2)/2^k)$ أو $[i/2^k, i/2^k + 1/2^{k-1})$ بالفترة الزمنية $1/2^{k-1}$. إنَّ الدالة $\psi(2^{k-1}x - i/2)$ في المعادلة ٢٠-٩ يكون لها القيمة 1 عندما تكــون قيــم x واقــعــة في $[i/2^k, (i + 1)/2^k)$ ويكــون لهــا القــيمــة 1- عنــدمــا تكــون قـــيــم x واقـــعـــة في النطــاق $[(i + 1)/2^k, (i + 2)/2^k)$. وهناك صيغة مكافئة للمعادلات ٢٠-٨ و٢٠-٩ وهي:

$$\varphi(2^k x - i) = \frac{1}{2}\left[\varphi\left(2^{k-1}x - \frac{i}{2}\right) + \psi\left(2^{k-1}x - \frac{i}{2}\right)\right] \qquad (٢٠\text{-}١٠)$$

$$\varphi(2^k x - i - 1) = \frac{1}{2}\left[\varphi\left(2^{k-1}x - \frac{i}{2}\right) - \psi\left(2^{k-1}x - \frac{i}{2}\right)\right] \qquad (٢٠\text{-}١١)$$

في الجانب الأيسر من المعادلات ٢٠-١٠ و٢٠-١١، ننظر إلى سجلات البيانات في الفترة الزمنية $1/2^k$ أو التكرار 2^k. في الجانب الأيمن من المعادلات ٢٠-١٠ و٢٠-١١، ننظر إلى سجلات البيانات في الفترة الزمنية الأكبر $1/2^{k-1}$ أو التكرار الأقل 2^{k-1}.

تسمح لنا المعادلتان ٢٠-١٠ و٢٠-١١ بتنفيذ تحويل المويجة لبيانات السلسة الزمنية أو بتمثيل دالتيهما في المعادلة ٢٠-٣ على شكل بيانات ذات تكرارات متنوعة كما هو موضح من خلال المثال ٢٠-١.

المثال ٢٠-١

قم بتنفيذ تحويل مويجة هار لبيانات السلسلة الزمنية التالية: ٠، ٢، ٠، ٢، ٦، ٨، ٦، ٨.
أولًا، نقوم بتمثيل بيانات السلسلة الزمنية باستخدام دالة القياس لمويجة هار:

$$f(x) = \sum_{i=0}^{2^k-1} a_i \varphi(2^k x - i)$$

$$\begin{aligned} f(x) = {} & 0\varphi(2^3x) + 2\varphi(2^3x - 1) \\ & + 0\varphi(2^3x - 2) + 2\varphi(2^3x - 3) \\ & + 6\varphi(2^3x - 4) + 8\varphi(2^3x - 5) \\ & + 6\varphi(2^3x - 6) + 8\varphi(2^3x - 7). \end{aligned}$$

ثم، نستخدم المعادلتين ٢٠-١٠ و٢٠-١١ لتحويل الدالة المذكورة آنفًا. عند تنفيذ تحويل المويجة للدالة المذكورة أعلاه، نستخدم $i=0$ ، و$i+1=1$ للزوج الأول من دوال القياس في الجانب الأيمن من الدالة المذكورة آنفًا، و$i=2$ و$i+1=3$ للزوج الثاني، و$i=4$ و$i+1=5$ للزوج الثالث، و$i=6$ و$i+1=7$ للزوج الرابع :

$$\begin{aligned} f(x) = {} & 0 \times \frac{1}{2}\left[\varphi\left(2^2x - \frac{0}{2}\right) + \psi\left(2^2x - \frac{0}{2}\right)\right] + 2 \times \frac{1}{2}\left[\varphi\left(2^{k-1}x - \frac{0}{2}\right) - \psi\left(2^{k-1}x - \frac{0}{2}\right)\right] \\ & + 0 \times \frac{1}{2}\left[\varphi\left(2^2x - \frac{2}{2}\right) + \psi\left(2^2x - \frac{2}{2}\right)\right] + 2 \times \frac{1}{2}\left[\varphi\left(2^{k-1}x - \frac{2}{2}\right) - \psi\left(2^{k-1}x - \frac{2}{2}\right)\right] \\ & + 6 \times \frac{1}{2}\left[\varphi\left(2^2x - \frac{4}{2}\right) + \psi\left(2^2x - \frac{4}{2}\right)\right] + 8 \times \frac{1}{2}\left[\varphi\left(2^{k-1}x - \frac{4}{2}\right) - \psi\left(2^{k-1}x - \frac{4}{2}\right)\right] \end{aligned}$$

$$+6\times\frac{1}{2}\left[\varphi\left(2^2x-\frac{6}{2}\right)+\psi\left(2^2x-\frac{6}{2}\right)\right]+8\times\frac{1}{2}\left[\varphi\left(2^{k-1}x-\frac{6}{2}\right)-\psi\left(2^{k-1}x-\frac{6}{2}\right)\right]$$

$$f(x)=0\times\frac{1}{2}[\varphi(2^2x)+\psi(2^2x)]+2\times\frac{1}{2}[\varphi(2^2x)-\psi(2^2x)]$$
$$+0\times\frac{1}{2}[\varphi(2^2x-1)+\psi(2^2x-1)]+2\times\frac{1}{2}[\varphi(2^2x-1)-\psi(2^2x-1)]$$
$$+6\times\frac{1}{2}[\varphi(2^2x-2)+\psi(2^2x-2)]+8\times\frac{1}{2}[\varphi(2^2x-2)-\psi(2^2x-2)]$$
$$+6\times\frac{1}{2}[\varphi(2^2x-3)+\psi(2^2x-3)]+8\times\frac{1}{2}[\varphi(2^2x-3)-\psi(2^2x-3)]$$

$$f(x)=\left(0\times\frac{1}{2}+2\times\frac{1}{2}\right)\varphi(2^2x)+\left(0\times\frac{1}{2}-2\times\frac{1}{2}\right)\psi(2^2x)$$
$$+\left(0\times\frac{1}{2}+2\times\frac{1}{2}\right)\varphi(2^2x-1)+\left(0\times\frac{1}{2}-2\times\frac{1}{2}\right)\psi(2^2x-1)$$
$$+\left(6\times\frac{1}{2}+8\times\frac{1}{2}\right)\varphi(2^2x-2)+\left(6\times\frac{1}{2}-8\times\frac{1}{2}\right)\psi(2^2x-2)$$
$$+\left(6\times\frac{1}{2}+8\times\frac{1}{2}\right)\varphi(2^2x-3)+\left(6\times\frac{1}{2}-8\times\frac{1}{2}\right)\psi(2^2x-3)$$

$$f(x)=\varphi(2^2x)-\psi(2^2x)$$
$$+\varphi(2^2x-1)-\psi(2^2x-1)$$
$$+7\varphi(2^2x-2)-1\psi(2^2x-2)$$
$$+7\varphi(2^2x-3)-1\psi(2^2x-3)$$

$$f(x)=\varphi(2^2x)+\varphi(2^2x-1)+7\varphi(2^2x-2)+7\varphi(2^2x-3)$$
$$-\psi(2^2x)-\psi(2^2x-1)-1\psi(2^2x-2)-1\psi(2^2x-3).$$

نقوم باستخدام المعادلتين ٢٠-١٠ و٢٠-١١ لتحويل السطر الأول من الدالة المذكورة آنفًا:

$$f(x) = \frac{1}{2}[\varphi(2^1x) + \psi(2^1x)] + \frac{1}{2}[\varphi(2^1x) - \psi(2^1x)]$$
$$+7 \times \frac{1}{2}[\varphi(2^1x-1) + \psi(2^1x-1)] + 7 \times \frac{1}{2}[\varphi(2^1x-1) - \psi(2^1x-1)]$$
$$-\psi(2^2x) - \psi(2^2x-1) - \psi(2^2x-2) - \psi(2^2x-3)$$

$$f(x) = \left(\frac{1}{2}+\frac{1}{2}\right)\varphi(2x) + \left(\frac{1}{2}+\frac{1}{2}\right)\psi(2x) + \left(\frac{7}{2}+\frac{7}{2}\right)\varphi(2x-1) + \left(\frac{7}{2}-\frac{7}{2}\right)\psi(2x-1)$$
$$-\psi(2^2x) - \psi(2^2x-1) - \psi(2^2x-2) - \psi(2^2x-3)$$

$$f(x) = \varphi(2x) - 7\varphi(2x-1)$$
$$+0\psi(2x) + 0\psi(2x-1)$$
$$-\psi(2^2x) - \psi(2^2x-1) - \psi(2^2x-2) - \psi(2^2x-3).$$

مرةً أخرى، نستخدم المعادلتين ٢٠-١٠ و٢٠-١١ لتحويل السطر الأول من الدالة المذكورة آنفًا:

$$f(x) = \frac{1}{2}[\varphi(2^{1-1}x) + \varphi(2^{1-1}x)] + 7 \times \frac{1}{2}[\varphi(2^{1-1}x) - \psi(2^{1-1}x)]$$
$$+0\psi(2x) + 0\psi(2x-1) - \psi(2^2x) - \psi(2^2x-1)$$
$$-\psi(2^2x-2) - \psi(2^2x-3)$$

$$f(x) = \left(\frac{1}{2}+\frac{7}{2}\right)\varphi(x) + \left(\frac{1}{2}+\frac{7}{2}\right)\psi(x)$$
$$+0\psi(2x) + 0\psi(2x-1) - \psi(2^2x) - \psi(2^2x-1)$$
$$-\psi(2^2x-2) - \psi(2^2x-3)$$

$$f(x) = 4\varphi(x) - 3\psi(x) + 0\psi(2x) + 0\psi(2x-1)$$
$$-\psi(2^2x) - \psi(2^2x-1) - \psi(2^2x-2) - \psi(2^2x-3). \qquad (١٢-٢٠)$$

تعطى الدالة في المعادلة ٢٠-١٢ النتيجة النهائية لتحويل مويجة هار. يوجد ثمانية حدود للدالة، كما أنَّ لعينة البيانات الأصلية ثمانية سجلات بيانات. الحد الأول، $4\varphi(x)$ يمثل دالة خطوة بارتفاع 4 لـ x في النطاق $[0,1)$ ويعطي متوسط سجلات البيانات الأصلية، 0، 2، 0، 2، 6، 8، 6، 8. الحد الثاني، $-3\psi(x)$، له دالة المويجة $\psi(x)$، وهو ما يمثل تغيير خطوة لقيمة الدالة من 1 إلى 1- أو تغيير خطوة بقيمة 2- كلما اتجهت قيم x من النصف الأول للنطاق $[0, ½)$ إلى النصف الثاني للنطاق $[½, 1)$. وبالتالي، فإن الحد الثاني، $-3\psi(x)$، يكشف أن بيانات السلسلة الزمنية الأصلية لديها تغيير خطوة مقداره (3-) × (2-) = 6 من مجموعة النصف الأول لسجلات البيانات الأربعة إلى مجموعة النصف الثاني لسجلات البيانات الأربعة إذا كان متوسط سجلات البيانات الأربعة الأولى مساوياً 1 ومتوسط سجلات البيانات الأربعة الأخيرة مساوياً 7. الحد الثالث، $0\psi(2x)$، يمثل أن بيانات السلسلة الزمنية الأصلية ليس لديها أي تغيير خطوة من سجلات البيانات الأولى والثانية إلى سجلات البيانات الثالثة والرابعة إذا كان متوسط سجلات البيانات الأولى والثانية مساوياً 1 ومتوسط سجلات البيانات الثالثة والرابعة مساوياً 1. الحد الرابع، $0\psi(2x-1)$، يمثل أن بيانات السلسلة الزمنية الأصلية ليس لديها أي تغيير خطوة من سجلات البيانات الخامسة والسادسة إلى سجلات البيانات السابعة والثامنة إذا بلغ متوسط سجلات البيانات الخامسة والسادسة 7 ومتوسط سجلات البيانات السابعة والثامنة 7. تكشف الحدود الخامسة، والسادسة، السابعة، والثامنة للدالة في المعادلة ٢٠-١٢، $-\psi(2^2x)$، $-\psi(2^2x-1)$، $-\psi(2^2x-2)$ و$-\psi(2^2x-3)$ أنَّ بيانات السلسلة الزمنية الأصلية لها تغيير خطوة (1-) × (2-) = 2 من سجل البيانات الأول بالقيمة صفر إلى سجل البيانات الثاني بالقيمة 2، وتغيير الخطوة (1-) × (2-) = 2 من سجل البيانات الثالث بالقيمة صفر إلى سجل البيانات الرابع بالقيمة 2، وتغيير الخطوة (1-) × (٢-) = 2 من سجل البيانات الخامس بالقيمة 6 إلى سجل البيانات السادس بالقيمة 8، وتغيير الخطوة (1-) × (2-) = 2 من سجل البيانات السابع بالقيمة 6 إلى سجل البيانات الثامن بالقيمة 8. وبالتالي، ينتج عن تحويل مويجة هار لثماني سجلات بيانات في بيانات السلسلة الزمنية الأصلية ثمانية حدود بمعامل دالة القياس $\varphi(x)$ كاشفًا عن متوسط البيانات الأصلية، ومعامل دالة المويجة $\psi(x)$ كاشفًا عن تغيير الخطوة في

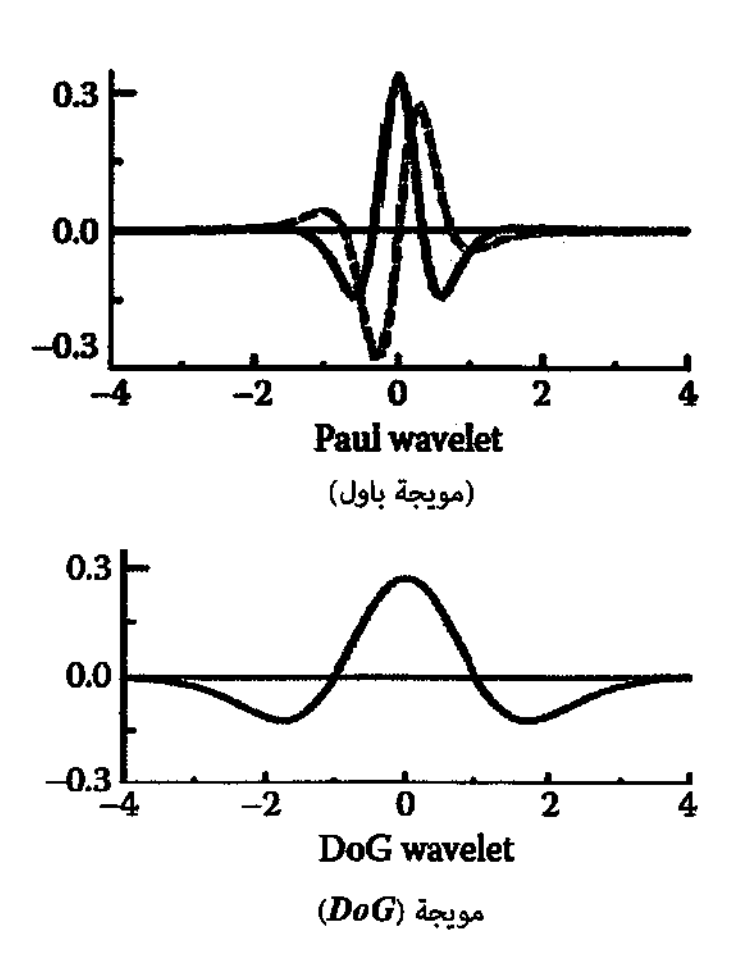
البيانات الأصلية بأقل تكرار من مجموعة النصف الأول لسجلات البيانات الأربع إلى مجموعة النصف الثاني لسجلات البيانات الأربع، وتكشف معاملات دالتي المويجات $\psi(2x)$ و$\psi(2x-1)$ عن تغييرات الخطوة في البيانات الأصلية عند أعلى تكرار لكل سجلي بيانات، وتكشف معاملات دالة المويجات $\psi(2^2x)$ و$\psi(2^2x-1)$، $\psi(2^2x-2)$ و$\psi(2^2x-3)$ تغييرات الخطوة في البيانات الأصلية عند أعلى تكرار لكل سجل بيانات .

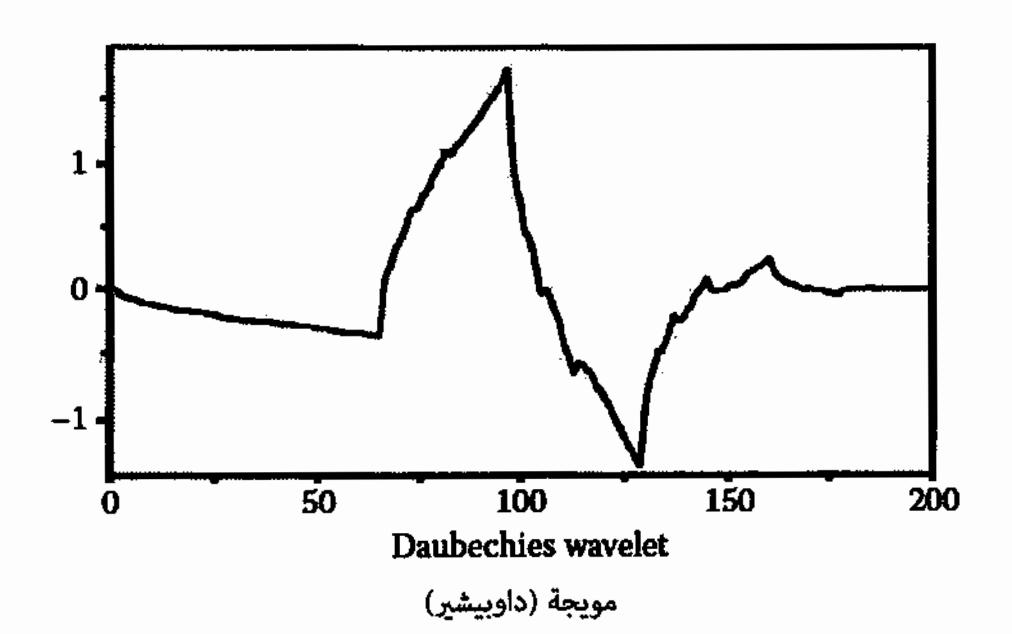
وبالتالي، فإن تحويل مويجة هار لبيانات السلسلة الزمنية يسمح لنا بتحول بيانات السلسلة الزمنية إلى البيانات في مجال التكرار الزمني، ورصد خصائص نمط بيانات المويجة (على سبيل المثال، تغيير الخطوة لمويجة هار) في مجال التكرار الزمني. على سبيل المثال، يكشف تحويل مويجة بيانات السلسلة الزمنية 0، 2، 0، 2، 6، 8، 6، 8 في المعادلة ٢٠-١٢ عن أنَّ البيانات لديها المتوسط 4، وزيادة قدرها 6 في الخطوة في أربعة سجلات بيانات (عند أدنى تكرار لتغيير الخطوة)، وليس هناك أي تغيير خطوة عند كل سجلي بيانات (عند التكرار المتوسط لتغيير الخطوة)، وزيادة قدرها 2 في الخطوة عند كل سجل بيانات (عند أعلى تكرار لتغيير الخطوة). بالإضافة إلى مويجة هار التي تلتقط نمط البيانات لتغيير الخطوة، فهناك العديد من أشكال المويجات الأخرى، على سبيل المثال، مويجة باول (*Paul wavelet*)، مويجة اشتقاق مويجة قوسشيان (*DoG*)، ومويجة داوبيشيز (*Doubechtes wavelet*)، ومويجة مورليت (*Morlet wavelet*) كما هو موضح في الشكل ٢٠-٣، والتي تلتقط أنواع أخرى من أنماط البيانات. يتم تطوير العديد من أشكال المويجات بحيث يمكن اختيار شكل المويجة المناسبة لإعطاء توافق قريب لنمط البيانات لبيانات السلسلة الزمنية. على سبيل المثال، يمكن استخدام مويجة داوبيشيز (*Daubechies, 1990*) لإجراء تحويل المويجة لبيانات السلسلة الزمنية التي تظهر نمط بيانات بزيادة خطية أو نقصان خطي. أما مويجة باول، ومويجة اشتقاق مويجة قوسشيان، فيمكن استخدامهما لبيانات السلسلة الزمنية التي تظهر أنماط بيانات مثل الموجة (*Wave- Like*).

الشكل (٢٠-٣)

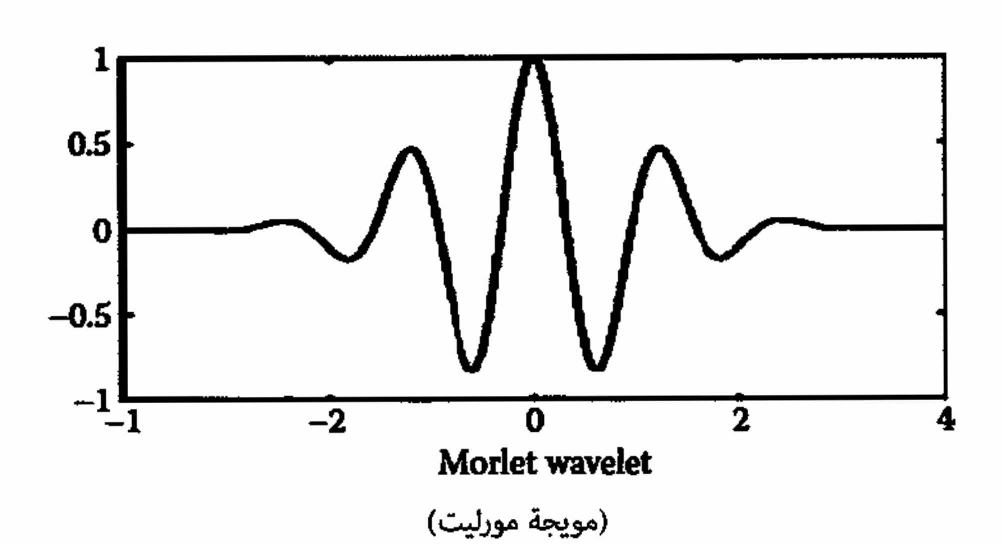
توضيح بياني لمويجة باول، ومويجة (DoG) اشتقاق مويجة قوسشيان، ومويجة داوبيشيز، ومويجة مورليت. (يي، إن، نظم الحاسوب والشبكة الآمنة: النمذجة والتحليل والتصميم، ٢٠٠٨، الشكل ١١,٢، ص ٢٠٠ حقوق الطبع والنشر لشركة وايلي في سي اتش فيرلاغ وشركاه المحدودة)

(Ye, N., Secure Computer and Network Systems:Modeling, Analysis - and Design, 2008, Figure 11.2, p. 200. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission)





مويجة (داوبيشير)



(مويجة مورليت)

٢٠-٣ إعادة بناء السلسلة الزمنية الزمن من معاملات المويجة (Reconstruction of Time Series Data from Wavelet Coefficients):

المعادلتان ٢٠-٨ و٢٠-٩، والتي يتم إعادة كتابتهما أدناه، يمكن استخدامهما لإعادة بناء بيانات السلسلة الزمنية من معاملات المويجة :

$$\varphi\left(2^{k-1}x-\frac{i}{2}\right)=\varphi(2^kx-i)+\varphi(2^kx-i-1)$$

$$\psi\left(2^{k-1}x-\frac{i}{2}\right)=\varphi(2^kx-i)-\varphi(2^kx-i-1).$$

المثال ٢٠-٢:

قم بإعادة بناء بيانات السلسلة الزمنية من معاملات المويجة في المعادلة ٢٠-١٢، والتي يتم تكرارها أدناه :

$$\begin{aligned}f(x)=&\;4\varphi(x)\\&-3\psi(x)\\&+0\psi(2x)+0\psi(2x-1)\\&-\psi(2^2x)-\psi(2^2x-1)-\psi(2^2x-2)-\psi(2^2x-3)\end{aligned}$$

$$\begin{aligned}f(x)=&\;4\times[\varphi(2^1x)+\varphi(2^1x-1)]\\&-3\times[\varphi(2^1x)-\varphi(2^1x-1)]\\&+0\times[\varphi(2^2x)-\varphi(2^2x-1)]+0\times[\varphi(2^2x-2)-\varphi(2^2x-3)]\\&-[\varphi(2^3x)-\varphi(2^3x-1)]-[\varphi(2^3x-2)-\varphi(2^3x-3)]-[\varphi(2^3x-4)-\varphi(2^3x-5)]\\&-[\varphi(2^3x-6)-\varphi(2^3x-7)]\end{aligned}$$

$$\begin{aligned}f(x)=&\;\varphi(2x)+7\varphi(2x-1)\\&-\varphi(2^3x)+\varphi(2^3x-1)-\varphi(2^3x-2)+\varphi(2^3x-3)-\varphi(2^3x-4)\\&+\varphi(2^3x-5)-\varphi(2^3x-6)+\varphi(2^3x-7)\end{aligned}$$

$$f(x) = [\varphi(2^2x) + \varphi(2^2x-1)] + 7 \times [\varphi(2^2x-2) + \varphi(2^2x-3)]$$
$$-\varphi(2^3x) + \varphi(2^3x-1) - \varphi(2^3x-2) + \varphi(2^3x-3) - \varphi(2^3x-4) + \varphi(2^3x-5)$$
$$-\varphi\left(2^3x-6\right) + \varphi\left(2^3x-7\right)$$

$$f(x) = \varphi(2^2x) + \varphi(2^2x-1) + 7\varphi(2^2x-2) + 7\varphi(2^2x-3)$$
$$-\varphi(2^3x) + \varphi(2^3x-1) - \varphi(2^3x-2) + \varphi(2^3x-3) - \varphi(2^3x-4) + \varphi(2^3x-5)$$
$$-\varphi(2^3x-6) + \varphi(2^3x-7)$$

$$f(x) = \left[\varphi(2^3x) + \varphi(2^3x-1)\right] + \left[\varphi(2^3x-2) + \varphi(2^3x-3)\right]$$
$$+7 \times [\varphi(2^3x-4) + \varphi(2^3x-5)] + 7[\varphi(2^3x-6) + \varphi(2^3x-7)]$$
$$-\varphi(2^3x) + \varphi(2^3x-1) - \varphi(2^3x-2) + \varphi(2^3x-3) - \varphi(2^3x-4)$$
$$+\varphi(2^3x-5) - \varphi(2^3x-6) + \varphi(2^3x-7)$$

$$f(x) = 0\varphi(2^3x) + 2\varphi(2^3x-1)$$
$$+0\varphi(2^3x-2) + 2\varphi(2^3x-3)$$
$$+6\varphi(2^3x-4) + 8\varphi(2^3x-5)$$
$$+6\varphi(2^3x-6) + 8\varphi(2^3x-7).$$

عند أخذ معاملات دالات القياس في الجانب الأيمن من المعادلة الأخيرة، فإنه يعطينا العينة الأصلية لبيانات سلاسل الزمن، 0، 2، 0، 2، 6، 8، 6، 8.

٢٠-٤ البرمجيات والتطبيقات (Software and Applications):

يتم دعم تحليل المويجة في حزم البرمجيات بما في ذلك برنامج ستاتيستيكا *STATISTIC* *(www.statistica.com)* وبرنامج ماتلاب *MATLAB* *(www.matworks.com)* كما نوقش في الجزء ٢٠-٢، يمكن تطبيق تحول المويجة للكشف عن خصائص أنماط بيانات معينة في مجال تكرار زمني. على سبيل المثال، عن طريق فحص موقع الزمن وتكرار معامل مويجة هار بالحجم الأكبر، تمّ الكشف عن حدوث أكبر صعود لمؤشر بورصة نيويورك لفترة ٦ سنوات من العام ١٩٨١-١٩٨٧ من أول ٣ سنوات إلى الثلاث سنوات التالية (*Boggess and Narcowich, 2001*). يمكن العثور على تطبيق مويجة هار، وپاول، ومويجة اشتقاق مويجة قوسشيان، ومويجة داوبيشيز، ومويجة مورليت .لبيانات الحاسوب والشبكات في يي (*Ye, 2008; Chapter 11*).

يُعتبر تحويل المويجة مفيدًا أيضًا لكثير من الأنواع الأخرى من التطبيقات، بما في ذلك خفض الضوضاء وتصفيتها، وضغط البيانات، والكشف عن الحافة (*Boggess and Narcowich, 2001*). وعادةً ما يتم القيام بخفض الضوضاء وتصفيتها عن طريق إسناد القيمة صفر لمعاملات المويجة في نطاق تكرار معين، والذي يؤخَذ في الاعتبار لتمييز الضوضاء في بيئة معينة (على سبيل المثال، أعلى تكرار للضوضاء البيضاء أو نطاق معين من التكرارات للضوضاء المتولدة آلياً في قُمْرَة قيادة طائرة إذا كان صوت الطيار هو محل الاهتمام). ثم يتم استخدام معاملات المويجة تلك جنبًا إلى جنب مع غيرها من معاملات المويجة الثابتة لإعادة بناء الإشارة بعد إزالة الضوضاء. وعادة ما يتم ضغط البيانات (*data compression*) من خلال الإبقاء على معاملات المويجة ذات المقدار الكبير أو معاملات المويجة عند بعض التكرارات التي تُعتبَر أنها تمثل الإشارة. يتم استخدام معاملات المويجة هذه وغيرها من معاملات المويجة الأخرى ذات القيمة صفر لإعادة بناء بيانات الإشارة. إذا تمّ نقل بيانات الإشارة من مكان إلى مكان آخر، وكلا المكانين يعرفان التكرارات المعطاة التي تحتوي على الإشارة، فهناك مجموعة صغيرة فقط من معاملات المويجة في التكرارات المعطاة تحتاج إلى أن تنتقل لتحقيق ضغط البيانات. يُعتبر الكشف عن الحافة (*edge detection*) بأنه البحث عن أكبر معاملات للمويجة واستخدام مواقع زمنهم وتكراراتهم في الكشف عن أكبر تغيير (تغييرات) أو انقطاعات في البيانات (على سبيل المثال، حافة حادة بين ظل خفيف إلى ظل داكن في صورة لكشف جسم ما كشخص في ردهة) .

التمارين (Exercises):

٢٠-١ قم بتنفيذ تحويل مويجة هار لبيانات السلسلة الزمنية 2.5، 0.5، 4.5، 2.5، 1-، 1، 2، 6 وشرح معنى كل معامل في نتيجة تحويل مويجة هار.

٢٠-٢ ينتج عن تحويل مويجة هار لبيانات سلسلة زمنية معينة معاملات المويجة التالية:

$$\begin{aligned} f(x) = {} & 2.25\varphi(x) \\ & +0.25\psi(x) \\ & -1\psi(2x) - 2\psi(2x-1) \\ & +\psi(2^2x) + \psi(2^2x-1) - \psi(2^2x-2) \\ & \quad - 2\psi(2^2x-3). \end{aligned}$$

قم بإعادة بناء بيانات السلسلة الزمنية الأصلية باستخدام هذه المعاملات.

٢٠-٣ بعد اسناد القيمة صفر للمعاملات التي تكون قيمها المطلقة أصغر من 1.5 في تحويل مويجة هار من التمرين ٢٠-٢، يكون لدينا معاملات المويجات التالية :

$$\begin{aligned} f(x) = {} & 2.25\varphi(x) \\ & +0\psi(x) \\ & +0\psi(2x) - 2\psi(2x-1) \\ & +0\psi(2^2x) + 0\psi(2^2x-1) + 0\psi(2^2x-2) \\ & \quad - 2\psi(2^2x-3). \end{aligned}$$

قم بإعادة بناء بيانات السلسلة الزمنية باستخدام هذه المعاملات.

المراجع

- Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, pp. 487–499.
- Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.
- Boggess, A. and Narcowich, F. J. 2001. *The First Course in Wavelets with Fourier Analysis*. Upper Saddle River, NJ: Prentice Hall.
- Box, G.E.P. and Jenkins, G. 1976. *Time Series Analysis: Forecasting and Control*. Oakland, CA: Holden-Day.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. *Classification and Regression Trees*. Boca Raton, FL: CRC Press.
- Bryc, W. 1995. *The Normal Distribution: Characterizations with Applications*. New York: Springer-Verlag.
- Burges, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2, 121–167.
- Chou, Y.-M., Mason, R. L., and Young, J. C. 1999. Power comparisons for a Hotelling's T2 statistic. *Communications of Statistical Simulation*, 28(4), 1031–1050.
- Daubechies, I. 1990. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5), 96–101.
- Davis, G. A. 2003. Bayesian reconstruction of traffic accidents. *Law, Probability and Risk*, 2(2), 69–89.
- Díez, F. J., Mira, J., Iturralde, E., and Zubillaga, S. 1997. DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, 10, 59–73.

- Emran, S. M. and Ye, N. 2002. Robustness of chi-square and Canberra techniques in detecting intrusions into information systems. *Quality and Reliability Engineering International*, 18(1), 19–28.
- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, U. M. Fayyad (eds.) *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR, AAAI Press, pp. 226–231.
- Everitt, B. S. 1979. A Monte Carlo investigation of the Robustness of Hotelling's oneand two-sample T2 tests. *Journal of American Statistical Association*, 74(365), 48–51.
- Frank, A. and Asuncion, A. 2010. UCI machine learning repository. http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.
- Hartigan, J. A. and Hartigan, P. M. 1985. The DIP test of unimodality. *The Annals of Statistics*, 13, 70–84.
- Jiang, X. and Cooper, G. F. 2010. A Bayesian spatio-temporal method for disease outbreak detection. *Journal of American Medical Informatics Association*, 17(4), 462–471.
- Johnson, R. A. and Wichern, D. W. 1998. *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall.
- Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59–69.
- Kruskal, J. B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1–27.
- Kruskal, J. B. 1964b. Non-metric multidimensional scaling: A numerical method. *Psychometrika*, 29(1), 115–129.
- Li, X. and Ye, N. 2001. Decision tree classifiers for computer intrusion detection. *Journal of Parallel and Distributed Computing Practices*, 4(2), 179–190.

- Li, X. and Ye, N. 2002. Grid- and dummy-cluster-based learning of normal and intrusive clusters for computer intrusion detection. *Quality and Reliability Engineering International*, 18(3), 231–242.
- Li, X. and Ye, N. 2005. A supervised clustering algorithm for mining normal and intrusive activity patterns in computer intrusion detection. *Knowledge and Information Systems*, 8(4), 498–509.
- Li, X. and Ye, N. 2006. A supervised clustering and classification algorithm for mining data with mixed variables. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 36(2), 396–406.
- Liu, Y. and Weisberg, R. H. 2005. Patterns of ocean current variability on the West Florida Shelf using the self-organizing map. *Journal of Geophysical Research*, 110, C06003, doi:10.1029/2004JC002786.
- Luceno, A. 1999. Average run lengths and run length probability distributions for Cuscore charts to control normal mean. *Computational Statistics & Data Analysis*, 32(2), 177–196.
- Mason, R. L., Champ, C. W., Tracy, N. D., Wierda, S. J., and Young, J. C. 1997a. Assessment of multivariate process control techniques. *Journal of Quality Technology*, 29(2), 140–143.
- Mason, R. L., Tracy, N. D., and Young, J. C. 1995. Decomposition of T2 for multivariate control chart interpretation. *Journal of Quality Technology*, 27(2), 99–108.
- Mason, R. L., Tracy, N. D., and Young, J. C. 1997b. A practical approach for interpreting multivariate T2 control chart signals. *Journal of Quality Technology*, 29(4), 396–406.
- Mason, R. L. and Young, J. C. 1999. Improving the sensitivity of the T2 statistic in multivariate process control. *Journal of Quality Technology*, 31(2), 155–164.
- Montgomery, D. 2001. *Introduction to Statistical Quality Control*, 4th edn. New York: Wiley.

- Montgomery, D. C. and Mastrangelo, C. M. 1991. Some statistical process control methods for autocorrelated data. *Journal of Quality Technologies*, 23(3), 179–193.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. *Applied Linear Statistical Models*. Chicago, IL: Irwin.
- Osuna, E., Freund, R., and Girosi, F. 1997. Training support vector machines: An application to face detection. In *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 130–136.
- Pourret, O., Naim, P., and Marcot, B. 2008. *Bayesian Networks: A Practical Guide to*
- *Applications*. Chichester, U.K.: Wiley.
- Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning*, 1, 81–106.
- Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.
- Rumelhart, D. E., McClelland, J. L., and the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. Cambridge, MA: The MIT Press.
- Russell, S., Binder, J., Koller, D., and Kanazawa, K. 1995. Local learning in probabilistic networks with hidden variables. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Montreal, Quebec, Canada, pp. 1146–1162.
- Ryan, T. P. 1989. *Statistical Methods for Quality Improvement*. New York: John Wiley & Sons.
- Sung, K. and Poggio, T. 1998. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1), 39–51.

- Tan, P.-N., Steinbach, M., and Kumar, V. 2006. *Introduction to Data Mining*. Boston, MA: Pearson.
- Theodoridis, S. and Koutroumbas, K. 1999. *Pattern Recognition*. San Diego, CA: Academic Press.
- Vapnik, V. N. 1989. *Statistical Learning Theory*. New York: John Wiley & Sons.
- Vapnik, V. N. 2000. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.
- Vidakovic, B. 1999. *Statistical Modeling by Wavelets*. New York: John Wiley & Sons.
- Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13, 260–269.
- Witten, I. H., Frank, E., and Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann.
- Yaffe, R. and McGee, M. 2000. *Introduction to Time Series Analysis and Forecasting*. San Diego, CA: Academic Press.
- Ye, N. 1996. Self-adapting decision support for interactive fault diagnosis of manufacturing systems. *International Journal of Computer Integrated Manufacturing*, 9(5), 392–401.
- Ye, N. 1997. Objective and consistent analysis of group differences in knowledge representation. *International Journal of Cognitive Ergonomics*, 1(2), 169–187.
- Ye, N. 1998. The MDS-ANAVA technique for assessing knowledge representation differences between skill groups. *IEEE Transactions on Systems, Man and Cybernetics*, 28(5), 586–600.
- Ye, N. 2003, ed. *The Handbook of Data Mining*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Ye, N. 2008. *Secure Computer and Network Systems: Modeling, Analysis and Design*. London, U.K.: John Wiley & Sons.
- Ye, N., Borror, C., and Parmar, D. 2003. Scalable chi square distance versus conventional statistical distance for process monitoring with uncorrelated data variables. *Quality and Reliability Engineering International*, 19(6), 505–515.
- Ye, N., Borror, C., and Zhang, Y. 2002a. EWMA techniques for computer intrusion detection through anomalous changes in event intensity. *Quality and Reliability Engineering International*, 18(6), 443–451.
- Ye, N. and Chen, Q. 2001. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International*, 17(2), 105–112.
- Ye, N. and Chen, Q. 2003. Computer intrusion detection through EWMA for autocorrelated and uncorrelated data. *IEEE Transactions on Reliability*, 52(1), 73–82.
- Ye, N., Chen, Q., and Borror, C. 2004. EWMA forecast of normal system activity for computer intrusion detection. *IEEE Transactions on Reliability*, 53(4), 557–566.
- Ye, N., Ehiabor, T., and Zhang, Y. 2002c. First-order versus high-order stochastic models for computer intrusion detection. *Quality and Reliability Engineering International*, 18(3), 243–250.
- Ye, N., Emran, S. M., Chen, Q., and Vilbert, S. 2002b. Multivariate statistical analysis of audit trails for host-based intrusion detection. *IEEE Transactions on Computers*, 51(7), 810–820.
- Ye, N. and Li, X. 2002. A scalable, incremental learning algorithm for classification problems. *Computers & Industrial Engineering Journal*, 43(4), 677–692.
- Ye, N., Li, X., Chen, Q., Emran, S. M., and Xu, M. 2001. Probabilistic techniques for intrusion detection based on computer audit data. *IEEE Transactions on Systems, Man, and Cybernetics*, 31(4), 266–274.

- Ye, N., Parmar, D., and Borror, C. M. 2006. A hybrid SPC method with the chi-square distance monitoring procedure for large-scale, complex process data. *Quality and Reliability Engineering International*, 22(4), 393–402.
- Ye, N. and Salvendy, G. 1991. Cognitive engineering based knowledge representation in neural networks. *Behaviour & Information Technology*, 10(5), 403–418.
- Ye, N. and Salvendy, G. 1994. Quantitative and qualitative differences between experts and novices in chunking computer software knowledge. *International Journal of Human-Computer Interaction*, 6(1), 105–118.
- Ye, N., Zhang, Y., and Borror, C. M. 2004b. Robustness of the Markov-chain model for cyber-attack detection. *IEEE Transactions on Reliability*, 53(1), 116–123.
- Ye, N. and Zhao, B. 1996. A hybrid intelligent system for fault diagnosis of advanced manufacturing system. *International Journal of Production Research*, 34(2), 555–576.
- Ye, N. and Zhao, B. 1997. Automatic setting of article format through neural networks. *International Journal of Human-Computer Interaction*, 9(1), 81–100.
- Ye, N., Zhao, B., and Salvendy, G. 1993. Neural-networks-aided fault diagnosis in supervisory control of advanced manufacturing systems. *International Journal of Advanced Manufacturing Technology*, 8, 200–209.
- Young, F. W. and Hamer, R. M. 1987. *Multidimensional Scaling: History, Theory, and Applications*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Glossary - قاموس المصطلحات

م	المصطلح العربي	المصطلح الإنجليزي
١	التعنقُد الهرمي المحتشد	Agglomerative hierarchical clustering
٢	المصفوفة الجبرية	Algebra matrix
٣	خوارزمية	Algorithm
٤	تحليل التباين	Analysis of variance
٥	تحليل تباين الزوايا	Angular analysis of variance
٦	شاذ	Anomaly
٧	خوارزمية أبريوري (الأسبقية)	Apriori algorithm
٨	الشبكة العصبية الصناعية	Artificial Neural Network (ANN)
٩	الاقتران	Association
١٠	أنماط الاقتران	Association patterns
١١	متغير الخاصية	Attribute variable
١٢	الارتباط الذاتي	Autocorrelation
١٣	ذاتي الانحدار	Autoregressive
١٤	نماذج المتوسط المتحرك ذاتي الانحدار	Autoregressive and moving average (ARMA) models
١٥	طريقة ترابط المتوسط	Average linkage method
١٦	طريقة التعلم بالتوالد الخلفي	Back-propagation learning method
١٧	مبدأ بيلمان	Bellman's principle
١٨	تحيز	Bias
١٩	التوزيع الثنائي النسق	Bimodal distribution
٢٠	تحويل بوكس-كوكس	Box–Cox transformation
٢١	متغير نوعي	Categorical variable
٢٢	المركز المتوسط	Centroid
٢٣	طريقة ترابط المركز المتوسط	Centroid linkage method
٢٤	إحصاءة مربع كاي	Chi-square statistic

م	المصطلح العربي	المصطلح الإنجليزي
٢٥	تصنيف	Classification
٢٦	عنقود	Cluster
٢٧	طريقة ترابط العناقيد	Cluster linkage method
٢٩	التعنقُد	Clustering
٢٩	تكلفة حاسوبية (معالجية، تخزينية، شبكية)	Computational cost
٣٠	الاحتمال المشروط	Conditional probability
٣١	مقياس الثقة	Confidence measure
٣٢	حد التحكم	Control limit
٣٣	ارتباط	Correlation
٣٤	تشابه جيب التمام (جتا)	Cosine similarity
٣٥	التغاير (التباين المشترك)	Covariance
٣٦	شرط أو معيار	Criterion
٣٧	مجموع تراكمي	Cumulative sum (CUSUM)
٣٨	الدرجة التراكمية	Cumulative score (CUSCORE)
٣٩	البيانات	Data
٤٠	تجانس البيانات	Data homogeneity
٤١	استكشاف أو تنقيب البيانات	Data Mining
٤٢	أنماط اختزال البيانات	Data reduction patterns
٤٣	مويجة داوبيشيز	Daubechies wavelet
٤٤	حد (حاجز) القرار	Decision threshold
٤٥	شجرة القرار	Decision tree
٤٦	رسم الدندروقرام الهرمي	Dendrogram
٤٧	دالة الكثافة	Density function
٤٨	اشتقاق مويجة قوسشيان	Derivative of Gaussian (DoG) wavelet
٤٩	المُحدد	Determinant

المصطلح الإنجليزي	المصطلح العربي	م
Deterministic trend	الاتجاه المحدَد	٥٠
Detrending	إعادة توجيه	٥١
Dilation effect	الأثر التمددي	٥٢
Dip test	اختبار أحادية النسق	٥٣
Directed, acyclic graph	الرسم البياني المفتوح والموجة	٥٤
Dissimilarity	اختلاف	٥٥
Edge detection	اكتشاف الحافة	٥٦
Eigenvalue	قيمة أيجن(القيمة الذاتية أو الجذر الكامن)	٥٧
Eigenvector	المتجه الذاتي	٥٨
Emission probability	احتمال الظهور	٥٩
Empirical risk of classification	مخاطرة التصنيف التجريبية	٦٠
Estimator	مقدِر	٦١
Euclidean distance	المسافة الإقليدية	٦٢
Expectation maximization	تضخيم التوقع	٦٣
Expected risk of classification	المخاطرة المتوقعة للتصنيف	٦٤
Exponentially weighted moving average (EWMA)	المتوسط المتحرك الموزون الأسّي	٦٥
False alarm rate	معدل الإنذار الخاطئ	٦٦
Feedforward ANNs	الشبكات العصبية الاصطناعية ذات التغذية الأمامية	٦٧
Gaussian Time Series	سلاسل قوسشيان الزمنية	٦٨
Gauss–Newton method	دالة قاوس-نيوتن	٦٩
Generalization	التعميم	٧٠
Gini index	مؤشر جيني	٧١
Goodness-of-fit	جودة المطابقة	٧٢
Gradient descent search	البحث الهابط المتدرج	٧٣
Graphical method	الأسلوب البياني	٧٤

المصطلح الإنجليزي	المصطلح العربي	م
Haar wavelet	مويجة هار	٧٥
Hamming distance	مسافة هامينغ	٧٦
Handwritten character recognition	تمييز الحروف المكتوبة بخط اليد	٧٧
Hard limit function	دالة الحد الثابت	٧٨
Hidden Markov models	نماذج ماركوف المخفية	٧٩
Histogram	المدرج التكراري	٨٠
Hit rate	معدل الزيارة الناجحة	٨١
Hotelling's T^2 control chart	مخطط التحكم لهوتلينق T^2	٨٢
Hotelling's T^2 statistic	إحصاءة هوتلينق T^2	٨٣
Hyperbolic tangent function	دالة الظل القطعي	٨٤
Identity matrix	المصفوفة المحايدة	٨٥
In-control process	عملية تحت السيطرة	٨٦
Independence of variables	استقلالية المتغيرات	٨٧
Individual difference scaling (INDSCALE)	قياس الفروقات الفردية	٨٨
Information	المعلومات	٨٩
Information Entropy	مقياس عشوائية المعلومات	٩٠
Interval variable	متغير الفترة	٩١
Inverse of a matrix	معكوس المصفوفة	٩٢
Joint probability	الاحتمال المشترك	٩٣
Karush–Kuhn–Tucker condition	شرط كاروش-كوهن-توكر	٩٤
Kernel function	دالة كيرنل	٩٥
K-nearest neighbor classifier	مُصنِّف أقرب k-مجاور	٩٦
Knowledge organization	تنظيم المعرفة	٩٧
Lagrange multiplier	مضاعف لاقرينج	٩٨
Least-squares method	طريقة المربعات الصغرى	٩٩

المصطلح الإنجليزي	المصطلح العربي	م
Lift measure	مقياس العَون	١٠٠
Linear classifier	مصنف خطي	١٠١
Linear function	الدالة الخطية	١٠٢
Linearly separable problem	مسألة قابلة للفصل خطيا	١٠٣
Log transformation	تحويل لوغاريتمي	١٠٤
Logistic regression model	نموذج الانحدار اللوجستي	١٠٥
Lower Control Limit (UCL)	حد التحكم الأدنى	١٠٦
Marginalization	تهميش	١٠٧
Markov chain	سلسلة ماركوف	١٠٨
Maximum a posterior (MAP) classification	تصنيف اللاحق (التالي) الأكبر	١٠٩
Maximum likelihood (ML) probability	احتمال الإمكان الأكبر	١١٠
Maximum likelihood method	طريقة الإمكان الأكبر	١١١
Maximum posterior probability	الاحتمال اللاحق الأكبر	١١٢
Mean shift	تحول المتوسط	١١٣
Measure of association	مقياس الاقتران	١١٤
Measure of data homogeneity	مقياس تجانس البيانات	١١٥
Minimum description length	طول الوصف الأصغر	١١٦
Minkowski distance	مسافة مينكوسكي	١١٧
Missing data	البيانات المفقودة	١١٨
Mode test	اختبار النسق	١١٩
Monotone regression algorithm	خوارزمية الانحدار الرتيبة	١٢٠
Monotonic tree of hierarchical clustering	التعنقُد الهرمي للشجرة الرتيبة	١٢١
Morlet wavelet	مويجة مورليت	١٢٢
Multidimensional scaling (MDS)	القياس المتعدد الأبعاد	١٢٣

المصطلح الإنجليزي	المصطلح العربي	م
Multilayer feedforward artificial neural	الشبكة العصبية الاصطناعية ذات التغذية الأمامية المتعددة الطبقات	١٢٤
Multimodal distribution	التوزيع المتعدد الأنساق	١٢٥
Multivariate control chart	مخطط التحكم المتعدد المتغيرات	١٢٦
Multivariate EWMA control chart	مخطط التحكم ذو المتوسط المتحرك الموزون الأسّي المتعدد المتغيرات	١٢٧
Multivariate statistics	إحصاءات المتغيرات المتعددة	١٢٨
Naïve Bayes Classifier	مصنف بييز البسيط	١٢٩
Natural language processing	معالجة اللغة الطبيعية	١٣٠
Natural logarithm transformation	التحويل اللوغاريتمي الطبيعي	١٣١
Neighborhood function	دالة المجاورة	١٣٢
Neural Network	الشبكة العصبية	١٣٣
Neuron	الخلية العصبية	١٣٤
Node	عُقدة	١٣٥
Noise reduction and filtering	اختزال الضوضاء وتصفيتها	١٣٦
Nominal variable	المتغير الإسمي	١٣٧
Nonbinary decision tree	شجرة القرار غير الثنائية	١٣٨
Nonlinear classifier	المُصنف غير الخطي	١٣٩
Nonlinear regression models	نماذج الانحدار غير الخطية	١٤٠
Nonlinearly separable problem	المسألة القابلة للانفصال بشكل غير خطي	١٤١
Non-monotonic tree of hierarchical clustering	التعنقُد الهرمي للشجرة غير الرتيبة	١٤٢
Nonstationarity	اللاسكون	١٤٣
Nonstationary time series	السلاسل الزمنية غير الساكنة	١٤٤
Normal distribution	التوزيع الطبيعي	١٤٥
Normal probability distribution	التوزيع الاحتمالي الطبيعي	١٤٦

المصطلح الإنجليزي	المصطلح العربي	م
Normalization method	دالة التطبيع	١٤٧
Normalized variable	المتغير المُطبع	١٤٨
Numeric variable	متغير رقمي	١٤٩
One-step ahead prediction model	نموذج التنبؤ بخطوة واحدة للأمام	١٥٠
Optimization problem	مشكلة التحسين	١٥١
Ordinal variable	المتغير الترتيبي	١٥٢
Orthogonal vector	المتجه المتعامد	١٥٣
Outlier	متطرف	١٥٤
Outlier and anomaly patterns	الأنماط المتطرفة والشاذة	١٥٥
Out-of-control process	عملية خارج السيطرة	١٥٦
Output unit	وحدة المخرجات	١٥٧
Over-fitted model	نموذج مفرط في المطابقة	١٥٨
Over-fitting	الإفراط في المطابقة	١٥٩
Parameter	مَعلَمة	١٦٠
Parameter estimation	تقدير المَعلَمة	١٦١
Partial autocorrelation function (PACF) coefficient	معامل دالة الارتباط الذاتي الجزئي	١٦٢
Pattern	نمط	١٦٣
Paul wavelet	مويجة باول	١٦٤
Perceptron	الشبكة العصبية الاصطناعية ذات التغذية الأمامية أحادية الطبقة	١٦٥
Polynomial function	دالة كثيرة الحدود	١٦٦
Positive definite matrix	المصفوفة المحددة الموجبة	١٦٧
Posterior probability	الاحتمال اللاحق	١٦٨
Prediction	تنبؤ	١٦٩
Principal component analysis	تحليل المكونات الرئيسية	١٧٠
Prior probability	الاحتمال السابق	١٧١

المصطلح الإنجليزي	المصطلح العربي	م
Probabilistic inference	الاستدلال الاحتمالي	١٧٢
Quadratic programming problem	مسألة برمجية تربيعية	١٧٣
Random fluctuation pattern	نمط التذبذب العشوائي	١٧٤
Random walk	السير العشوائي	١٧٥
Ratio variable	المتغير النسبي	١٧٦
Receiver Operating Curve (ROC)	منحنى التشغيل التشخيصي	١٧٧
Reconstruction of time series data	إعادة تشكيل بيانات السلسلة الزمنية	١٧٨
Recurrent ANNs	الشبكات العصبية الاصطناعية الدورية	١٧٩
Reduction	اختزال	١٨٠
Regression model	نموذج الانحدار	١٨١
Residual	المُتبقي	١٨٢
Scaling function	دالة القياس	١٨٣
Seasonable cycle	الدورة الموسمية	١٨٤
Self-Organizing Map (SOM)	خريطة التنظيم الذاتي	١٨٥
Sequential	تسلسلي	١٨٦
Sequential and temporal patterns	الأنماط الزمنية والتسلسلية	١٨٧
Shewhart control charts	مخطط شوارتز للتحكم	١٨٨
Shift effect	أثر التحول	١٨٩
Sigmoid function	الدالة السينية(على شكل حرف اس)	١٩٠
Sign function	دالة الإشارة	١٩١
Skewed distribution	التوزيع الملتوي	١٩٢
Skewness	الالتواء	١٩٣
Spectral decomposition of a matrix	التحلل الطيفي لمصفوفة	١٩٤
Speech recognition	التعرف على الكلام	١٩٥

المصطلح الإنجليزي	المصطلح العربي	م
Spike pattern	النمط المسماري	١٩٦
Split selection methods	دوال انتقاء الانفصال	١٩٧
State transition probability	احتمال تحول الحالة	١٩٨
Stationarity	السكون	١٩٩
Stationary time series	السلاسل الزمنية الساكنة	٢٠٠
Structural risk minimization principle	مبدأ تقليل المخاطر الهيكلية	٢٠١
Sum of squared errors (SSE)	مجموع الأخطاء التربيعية	٢٠٢
Supervised clustering	التعنقُد المراقب	٢٠٣
Support measure	مقياس الدعم	٢٠٤
Support vector machines (SVM)	الدعم الآلي المتجه	٢٠٥
Symmetric matrix	المصفوفة المتناظرة	٢٠٦
Target variable	متغير الهدف	٢٠٧
Temporal	زمني	٢٠٨
Tensor product	الضرب الممتد	٢٠٩
Test data	البيانات الاختبارية	٢١٠
Time series analysis	تحليل السلاسل الزمنية	٢١١
Training data	البيانات التدريبية أو الاستكشافية	٢١٢
Uniform distribution	التوزيع الموحد	٢١٣
Univariate	أحادي المتغير	٢١٤
Upper Control Limit (UCL)	حد التحكم الأعلى	٢١٥
Variance	تباين	٢١٦
Variance-covariance matrix	مصفوفة التباين-التغاير	٢١٧
VC dimension	بعد فابينك وتشرفونينكيس	٢١٨
Viterbi algorithm	خوارزمية فيترباي	٢١٩
Wavelet	مويجة	٢٢٠
Wavelet function	دالة المويجة	٢٢١

المترجم في سطور

الدكتور خالد بن ناصر آل حيان

المؤهل العلمي:

- حاصل على شهادة الدكتوراه في تخصص نظـم المعلومـات من جامعة جنـوب فلـوريدا بمدينة تامبا، ولاية فلوريدا، الولايات المتحدة الأمريكية في عام ١٤٣٤ هـ / ٢٠١٢ م.

العمل الحالي:

- مدير إدارة استشارات المعلومات والتقنية في معهد الإدارة العامة.

الأنشطة العلمية والعملية:

- له العديد من المؤلفات العلمية ما بين أوراق عمل علمية ومُترجَمات، إضافةً لهذا الكتاب، وتشمل على سبيل المثال:

م	النوع	المؤلف العلمي
١	ورقة عمل	Alhayyan,K., " **Participation in Information Markets Research: A New Conceptualization and Measurement**," Journal of Systemics, Cybernetics and Informatics (JSCI), Vol. 13 – No. 2 – Sep 2015, , pp. 68-76.
٢	ترجمة كتاب	ترجمة كتاب "**Social Science Research: Principles, Methods, and Practices**" – "بحوث العلوم الاجتماعية: المبادئ والمناهج والممارسات"، للمؤلف د. أنول باتشيرجي، سنة النشر ٢٠١٥م، دار اليازوري للنشر والتوزيع، ٤٢٧ صفحة.
٣	ورقة عمل	Alhayyan. K., Nuseibeh, H., "**Trends in the study of Cloud Computing: Observations and Research Gaps**", The 5th International Conference on Society and Information Technologies: ICSIT 2014, March 4-7 2014, Proceedings Vol. 1, pp. 38-43.

م	النوع	المؤلف العلمي
٤	ترجمة مقال	ترجمة مقال علمي بعنوان "الاتجاهات الخاصة بدراسة الإدارة العامة: ملاحظات تجريبية ونوعية من مجلة مراجعة الإدارة العامة ٢٠٠٠ ــ ٢٠٠٩م"، للمؤلفين: جوز سي إن. رادشيلدرز. كوانغ – هون لي، مجلة الإدارة العامة، المجلد رقم ٥٤، العدد ١، سنة النشر نوفمبر ٢٠١٣م.
٥	مراجعة ترجمة	مراجعة ترجمة مقال علمي بعنوان "تصميم نظم للتعلم الإلكتروني ذات وعي اجتماعي من خلال إدارة المعرفة"، للمؤلفين: ريشا شارما . هيما باناتي . بونام بيدي، ترجمة الدكتور/ عجلان بن محمد الشهري، مجلة الإدارة العامة، المجلد رقم ٥٣، العدد ٤، سنة النشر أغسطس ٢٠١٣م.
٦	ورقة عمل	Alhayyan,K., " **Cloud Computing: Better Ways to Control its Services,**" The 3rd International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2012, March 25th – 28th 2012, Proceedings Vol. 1, pp. 145-148.
٧	ورقة عمل	Alhayyan,K. , Bouayad, L., " **A Data Mining Method for the Medical Relationship between Diagnoses and Procedures – Vermont Hospital 2009,**" The 3rd International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2012, March 25th – 28th 2012, Proceedings Vol. 1, pp. 1-6.
٨	ورقة عمل	Alhayyan,K. , Collins, R., Jones, J. , Berndt, D., "**Economic Culture and Prediction Markets,**" Journal of Systemics, Cybernetics and Informatics (JSCI), Vol. 9 – No. 6 – Dec 2011, , pp. 69-74.

- يعمل مُحكّمًا ومُراجعًا للعديد من الأعمال العلمية والإدارية داخل المملكة العربية السعودية، كجامعة الملك سعود، ووزارة الإعلام، ومعهد الإدارة العامة، وهيئة الخبراء بمجلس الوزراء، وخارج المملكة العربية السعودية، كمؤتمرات WMSCI وIREPS.
- تصميم الحقائب التدريبية في معهد الإدارة العامة في مجال تقنية المعلومات.
- رئيس لجنة إعداد الخطة الإستراتيجية لتقنية المعلومات في معهد الإدارة العامة في عام ١٤٣٦/١٤٣٥هـ والتي يمتد تنفيذها إلى عام ١٤٤٠هـ
- منسق فريق (١٤٣٧/١٤٣٦ هـ)، ، في إعداد معايير اعتماد نشاط الاستشارات في معهد الإدارة العامة بالتنسيق مع الهيئة الوطنية للتقويم والاعتماد الأكاديمي في وزارة التعليم.
- مبرمج ومحلل تطبيقات برمجية في وزارة الدفاع والطيران والمفتشية العامة خلال الفترة من ١٩٩٠ إلى ١٩٩٧م، وكمبرمج ومحلل تطبيقات برمجية متعاون في دارة الملك عبدالعزيز عام ١٩٩٨م، وكمبرمج ومحلل تطبيقات برمجية متعاون في الاتحاد السعودي للفروسية عام ١٩٩٣م.

مراجع الترجمة في سطور

الدكتور صالح بن محمد السليم

المؤهل العلمي:

- حاصل على درجة الدكتوراه من جامعة واين ستيت بولاية ميشيغان، الولايات المتحدة الأمريكية، عام ٢٠٠١م في مجال علوم الحاسب (الذكاء الصناعي).

العمل الحالي:

- أستاذ مشارك في كلية علوم الحاسب والمعلومات، جامعة الملك سعود.

الأنشطة العلمية والعملية:

شغل العديد من المناصب منها القبول والتسجيل في جامعة شقراء، وشغل أيضاً منصب عميد تقنية المعلومات والتعليم الإلكتروني في جامعة شقراء كان يعمل سابقاً رئيساً لقسم تقنية المعلومات في الجامعة العربية المفتوحة، وقبل ذلك كان يعمل رئيساً لقسم تقنية الحاسب وعضو هيئة التدريس في الكلية التقنية بالرياض.

الاهتمامات البحثية تشمل التالي: الحساب التطويري، تصنيف النصوص، نئم تخطيط موارد المؤسسات، إدارة إجراءات الأعمال، التعليم الإلكتروني، والبرمجيات مفتوحة المصدر



تصميم وإخراج وطباعة الإدارة العامة للطباعة والنشر
بمعهد الإدارة العامة - ١٤٣٧هـ


MOHAMED KHATAB

هذا الكتاب

"يقدم هذا الكتاب تغطية شاملة لأهم الموضوعات في مجال استكشاف البيانات. ويستطيع القارئ الحصول على نظرة شاملة في استكشاف البيانات بما في ذلك المفاهيم الأساسية، والمسائل المهمة في هذا المجال، والكيفية التي يتم بها معالجة هذه المسائل. يتم تقديم الكتاب بطريقة تمكن القارئ الذي ليس لديه خلفية معرفية كافية في استكشاف البيانات، من الفهم بيسر وسهولة. كما يُمكنُ للقارئ الاطلاع على العديد من الأشكال الرسومية والأمثلة البديهة في هذا الكتاب. وأجدُ نفسي مولعاً بهذه الأشكال والأمثلة لأنها تجعل من المفاهيم والخوارزميات الأكثر تعقيداً أكثر سهولة للفهم."

- زهينق زهاو (Zheng Zhao)، معهد ساس (SAS)، كاري، كارولاينا الشمالية، الولايات المتحدة الأمريكية

"يغطي هذا الكتاب بشكل كبير كل خوار زميات استكشاف البيانات الأساسية. كما أنه يغطي العديد من الموضوعات المفيدة والتي لا يتم التطرق لها في الكتب الأخرى الخاصة باستكشاف البيانات، مثل موضوعات مخططات التحكم أحادية المتغير ومخططات التحكم متعددة المتغيرات وتحليل المويجة. ويتميز الكتاب بتوظيفه لأمثلة مفصّلة توضح الاستخدام العملي لخوارزميات استكشاف البيانات. كما يستعرض الكتاب قائمة من الحزم البرمجية الملائمة لتطبيق معظم الخوارزميات التي تم تغطيتها في الكتاب. ويُعتبَر هذا التوظيف للأمثلة والحزم البرمجية مفيداً إلى حدّ كبير لممارسي استكشاف البيانات. أوصي بقراءة هذا الكتاب لأي فرد مهتم باستكشاف البيانات."

- جيبينق يي (Jieping Ye)، جامعة أريزونا الحكومية، تيمبي، أريزونا، الولايات المتحدة الأمريكية

تُتيح التقنيات الحديثة جمع كميات هائلة من البيانات في العديد من المجالات. وبالرغم من ذلك فإن السرعة في اكتشاف معلومات ومعرفة مفيدة من هذه البيانات أقلُ بكثير من السرعة في جمع تلك البيانات. يستعرض كتاب، استكشاف البيانات: نظريات وخوارزميات وأمثلة، ويشرح مجموعة شاملة من خوارزميات استكشاف البيانات مستقاة من مجالات متنوعة لاستكشاف البيانات. كما يستعرض الكتاب التبريرات النظرية والتفاصيل الإجرائية لخوارزميات استكشاف البيانات، بما في ذلك تلك الخوارزميات الشائعة في الدراسات العلمية السابقة وتلك الخوارزميات ذات الصعوبة الكبيرة في الفهم، باستخدام عدة مجموعات من البيانات الصغيرة لشرح وتتبع خطوات تنفيذ كل خوارزمية.



تصميم وإخراج وطباعة الإدارة العامة للطباعة والنشر - معهد الإدارة العامة ١٤٣٧هـ

9960142444